THE UNIVERSITY OF ALBERTA

A FACTOR ANALYTIC ITEM SELECTION PROCEDURE

BY

MERLIN WALTER WAHLSTROM

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

APRIL, 1967

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "A Factor Analytic Item-Selection Procedure" submitted by Merlin Walter Wahlstrom in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

The problem investigated was concerned with developing a theoretical basis for an item-selection algorithm using factor analystic methods. After a test has been administered to a group of subjects and criteria variables are obtained, an item-criterion intercorrelation matrix is calculated. The intercorrelation matrix is then factor analyzed to determine the factors that define the item and criterion space. A rotation of the resulting orthogonal factor structure is applied to provide a final solution that has simple structure properties. Each factor is then assigned a relative weight by the test constructor to determine the position of a hypothetical goal vector in the factor space. The goal vector defines the desired characteristics of the test to be constructed.

Initially, the two items having the largest correlations with the goal vector are selected. A composite vector is formed by calculating the centroid of the two selected items. Prior to selecting the next item, the characteristics for the item to be selected are defined so that the goal vector and the composite vector are nearly collinear. Each additional item selected has properties that are the best approximation, within limits of the items available for selection, for producing collinearity. When all items for constructing a test have been selected, the centroid of the k selected items then determines the location of the composite vector.

An estimate of the constructed test's validity is given by

the correlation between the composite vector and the goal vector. Reliability is defined as the proportion of variance accounted for by a test vector. Two reliabilities can easily be obtained by considering the item clustering about either the goal vector or the composite vector. Since items are selected according to the goal vector's characteristics, a meaningful value would be in terms of item projections on the goal test. However, a truly internal consistency estimate of reliability is obtained by using the composite test vector which has as co-ordinates the centroid of the selected items.

The procedure for selecting items is not intended to replace existing item analysis methods but rather extends the analytic approach of the test constructor. In the proposed method, primary consideration has been given to meaningfulness and practicality of application. With electronic machines to handle the major part of selecting items, effort on tedious nonprofitable tasks should be reduced.

ACKNOWLEDGMENTS

The writer wishes to acknowledge gratefully the assistance of the members of his supervisory and candidacy committees in the preparation of this dissertation. In particular, Dr. S. M. Hunka, chairman of the committee, provided guidance without which the writer could not have succeeded.

The constructive suggestions of Dr. F. J. Boersma, Dr. D. P. McLeod, Dr. L. D. Nelson, and Dr. K. Smillie are most appreciated.

Finally, the author is grateful to the University of Alberta for providing the financial assistance which allowed him to pursue graduate studies and complete this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In recent years there has been considerable interest in objectifying measurement and evaluation procedures (Greene, Jorgenson and Gerberich, 1954; Thorndike and Hagen, 1955; Helmstadter, 1964). Increasing emphasis has been given to multiple-choice tests that can be machined scored.

It is commonly accepted by test experts that multiple-choice tests are the most highly regarded and widely used form of objective test item. "Almost any understanding or ability that can be tested by means of any other item form . . . can also be tested by means of multiple-choice test items" (Ebel, 1965, p. 149). Although there has been much criticism regarding the use of multiple-choice tests (Hoffman, 1962), the critics seldom seriously attempt to make a good case for a better way of measuring educational achievement.

While the mechanics of handling test administration and test scoring have readily been adapted to our "modern era" by the use of optical scanners such as the DIGITEK, MRC, and IBM machines, few applications of mathematical algorithms have been made for selecting items to construct a test. A review of the literature has revealed several methods (Wherry and Gaylord, 1946; Webster, 1956; Elfving, Sitgreaves and Solomon, 1959; Flowers, 1965) for selecting items but few test constructors have published descriptions of practical applications of the procedures.

With the ordinary computing methods used by many researchers, the

labor of even approximate solutions proposed in some selection procedures makes the techniques impractical. With the development of high speed electronic computers, exact solutions to the problem may prove to be quite practical.

There is a prevalent need for an analytic item-selection procedure. Items have been written for all age and grade levels. Perhaps the most productive persons have been those teaching junior high school, senior high school and university courses. Many pools of items are already available. At the university level, published lists of items (Hilgard, 1962; Orleans, 1963) are available and many instructors have accumulated items throughout the years. Item analysis techniques have been used extensively by test constructors for selecting items (Davis, 1951; Nunnally, 1959; Adams and Torgerson, 1964). The procedure of using item analysis data becomes tedious and time consuming when several item characteristics are evaluated simultaneously for many items. Analytic procedures are required that use a greater proportion of the statistical information available. Use of approximate solutions should yield to exact techniques.

The procedures for selecting items to construct a test seem to parallel other psychometric developments. In factor analytic theory, the rotation of axes has been a perennial problem. At first rotations were done by hand with simple mechanical aids. A theory for rotation was presented by Thurstone (1947) and his followers. Later equations were derived that were subsequently used as a criterion for rotation (Carroll, 1953; Newhaus and Wrigley, 1954; Kaiser, 1958; Saunders, 1960).

As a result of the increased computational effort required for even a small problem, the use of computers was introduced to make the procedure feasible. Some analytic procedures are being used in test construction but in these procedures considerable information is not used. Tedious procedures are still in use. Slowly there is evolving statistical methodology that is finding application because electronic machines have reduced the necessary hand computations to a minimal level. The procedure suggested by the writer centers upon an analytic method for selecting items from a pool of items. Although the proposed method for selecting items involves a considerable amount of calculation, this does not pose a problem. "As the computer revolution continues in psychometrics, we can expect objective algorithmic methods to become the rule rather than the exception" (Green, 1966, p. 444). The need for an automatic analytic item-selection technique is evident in our educational system where we continually construct new tests and modify our old ones.

The proposed selection technique is to some extent flexible for individual users who require a test with specific characteristics. Within limits, a desired reliability and validity estimate of the final test can be obtained by selecting the appropriate items. This can be done automatically. Factors, from a factor analysis of the items and criteria, are used as a basis to select the items to construct a test. Consideration has been given to developing a practical and objective method for selecting the "best subset" of items from a given pool of items even if hypothetical criteria must be devised.

The proposed selection method is perhaps better designated as the revision of a test with $\underline{n}$ items into a test with $\underline{k}$ items ($k<n$) since

the 'item pool' idea generally does not require that all items have been given to the same group or in the same test. Thus, a restriction is placed upon the definition of an item pool as used in this study in that only items that have been administered to the same group of subjects and within the same test format will be used to form the item pool.

The solution is amenable to computation by an electronic computer. A number of factors complicate the item selection problem not least of which are test reliability, test validity, and the resulting test score distribution.

CHAPTER II

GENERAL PROBLEM

Many objective items are now available to form pools of items. However, because of the varying nature of the presently available items, it is not sufficient to merely collect items and form pools of items. Prior to the inclusion of an item in a pool each item should be inspected by a sophisticated judge to determine whether obvious flaws are apparent in the item construction (Ebel, 1965). Common characteristics to be investigated are the precision with which the problem and solution are stated and the appropriateness of the item being made a part of an item universe. The next step, following the administration of the items to a group of subjects, is a statistical analysis of the items. An item analysis will provide further information about the quality of an item. Although consideration must be given to the statistical and authoritative judgements in evaluating items, the collection of items should be carefully constructed in such a way as to sample broadly the desired content and educational objectives of the pool. The content and educational objectives should be described prior to the writing and/or selection of the items for the item pool.

At present the main procedure for selecting items to construct a test is item analysis. While several statistical item characteristics are obtained from an item analysis, the technique does not readily lend itself to provide an answer as to which is the "best" item, "second best" item and so on. What is required is an analytic method for selecting

the best subset of items from the available pools of items.

While item analysis procedures can be used to evaluate the statistical acceptability of an item, the item parameters obtained are not in a form that permits a test constructor, in many situations, to decide readily which item is the "best" item of a pair of items. The problem becomes extremely more complicated when there are several items of which only the few "best" items are to be selected. It is assumed, for the purpose of the present discussion, that the relative importance of the content and learning objectives in the area under examination have been considered in relation to each item that was subjected to an item analysis. In addition, the most general meaning has been intended in the use of the term item analysis since "there is no one type of item analysis data that is best under all circumstances" (Davis, 1951, p. 297).

A procedure commonly used in test construction is that of writing each item on a card and then adding the relevant item analysis datum as it is collected. In this way, a pool of items is obtained. When a new test is to be constructed, the test constructor may select items from the pool that are acceptable to him. The items are then used to construct a test. Although the above procedure has merit, two major problems exist. The first problem is the difficulty of interpreting the relevance and significance of the item analysis parameters. If the items have been administered in several different tests to many groups of subjects, the item parameters are not directly comparable. A second problem is the dimensionality of the test. Without a statistical analysis of the selected items, no knowledge is available

regarding the nature and the number of dimensions the test will have.

An analytic technique is required that will select the "best" item, "second best" item and so on. The procedure suggested here, which utilizes factor analytic theory, for selecting items is not intended to replace existing item analysis methods but rather to extend the analytic approach of the test constructor. Item analysis data must be inspected before an item becomes part of a pool of items. After a pool of items is formed, the proposed algorithm may be used to select items according to the specific criterion established by the test constructor.

The summary statistical information available on a particular item embedded in an adequately defined data matrix, coupled with the power of an electronic computer, should result in a superior procedure for item selection than item analysis procedures. It is difficult to determine the relative importance of the item validity coefficient and the discrimination index when comparing many items. The added information of dimensionality and location of an item vector in an item and criterion space provided by factor analysis, is lacking in item analysis methods. Therefore more information is available through the factor analytic technique than through standard item analysis.

A primary consideration is meaningfulness and practicality. Any method should require minimal effort on tedious nonprofitable tasks. This can be done by using electronic equipment and suitable numerical procedures. By providing a general solution, variations desired by the test user may be developed by specifying required parameters.

In the proposed method, the onus is on the test constructor to provide estimates of certain desirable test characteristics. A test is constructed by selecting items to form a test which is dependent upon the established tolerance limits set by the user and the nature of the item pool from which the items are to be selected. The constructed test is to be an acceptable approximation of a postulated hypothetical test. Each user of the proposed technique should be able to construct his own pool of items which satisfy certain conditions deemed necessary. Such flexibility should be allowed within an analytic system. However, while the selection of items can conceivably be carried out by the proposed algorithm on an electronic computer and thus result in a test being constructed that has been "untouched by human hands", it is undesirable to have a test constructed that has been "untouched by human minds". Decisions regarding criteria, type of item, length of a test, and final test characteristics remain the jurisdiction of an "informed" test constructor.

Lumsden (1961) prepared a general survey of the construction of unidimensional tests. Greatest emphasis was "deliberately placed on item selection rationale since this topic appears to have been relatively neglected in the literature of the problem" (Lumsden, 1961, p. 130). The general conclusion advanced was that only factor analysis provides a rational procedure for item selection in the construction of unidimensional tests.

The theoretical development of measurement theory has primarily been concerned with unidimensional tests. It is the writer's contention that most achievement tests and ability tests are multidimensional. If

this is the case, we should be concerned with selecting items based upon an awareness of the multidimensional nature of the predictor variables and the criteria. The selection of items, to construct a test from a pool of items, is not generally done by random selection procedures. Consideration of difficulty and a discrimination index for each item enters into the decision as to the selection or the rejection of an item. In addition, each dimension of a multidimensional test should receive recognition in the item selection process. A special case of selecting items from a multidimensional space occurs when the item and criterion space is defined as unidimensional.

As in most methods of test construction, the "criterion problem" must be considered in the proposed item selection procedure. In an effort to structure the criterion problem and some possible solutions, Astin (1964) presented a paper concerned with clarifying certain issues regarding criterion measures and their use in educational and psychological research.

Although Astin deals with the problem of multiple criterion elements, the multidimensionality of criteria and the uniqueness of criterion measures found in some studies (Ryans, 1966; Kelly, 1966) suggests that greater effort should be made to produce an acceptable procedure for dealing with multiple criteria. Gulliksen (1950 b) has suggested that the most information about the criterion is available when a comprehensive matrix of intercorrelations including both predictor and criterion variables is utilized. Although Astin is critical of Gulliksen's recommendation in that it "involves circular reasoning or, at best, misuse of terms" (Astin, 1964, p. 812), Rozeboom (1966)

appears to be in agreement with Gulliksen. Rozeboom suggests that

> the concept of "validity" (can be) generalized from predicting a single criterion to predicting within a space, S_Y, of criterion variables . . . which makes clear that the concept of test penetrance into criterion space is a natural generalization of single-criterion validity theory (1966, pp. 442 - 443).

The writer is in agreement with Rozeboom and has incorporated the suggested "space concept" of criterion variables into the item-selection method.

Thorndike (1949), Gulliksen (1950 a) and Astin (1964) are in agreement that, in the final analysis, a judge or panel of "experts" must decide on rational grounds how relevant each element is to the conceptual criterion. The relationship between qualitative and quantitative decisions is especially relevant to the "criterion problem". Gulliksen aptly summarizes the situation in saying that

> mathematical procedures are appropriately used when they serve to guide thought. If an attempt is made to utilize such routines as a substitute for thought, we may unwittingly arrive at and accept absurd conclusions (1950 a, p. 351).

Since in the method proposed an hypothetical test must be specified, in the form of weights assigned to each criterion factor, each constructed test should be more acceptable and meaningful to the test user than existing tests prepared by other methods. As items are added to the item pool, the factor analysis of the included items will reveal any change in the nature of the pool. Thus, some control can be maintained over the inclusion of items similar and/or different from those in the existing pool. It is thus at the discretion of the user whether the item pool should remain the same, in a factor space sense, or be changed in a specified manner. Therefore, with knowledge of the

factor pattern of the item and criterion space, as well as the associated dimensions, more information is being used in selecting items. This should result in improved test construction practices.

An immediate reaction can occur from test specialists regarding the assumption of homogeneity of items or unidimensionality. One solution to the problem of working with multidimensional tests is to consider each test as if it was a sub-test of a test battery. However, test constructors should be made aware of the fact that many tests are multidimensional while being considered as unidimensional. The number of dimensions that an item and criterion space occupy is defined as being the number of orthogonal vectors required to span the space as defined by the item and criterion intercorrelation matrix.

Initially the proposed procedure uses the intercorrelation matrix of items and criteria as basic data. The criteria are not necessary but are desirable in providing for introducing an item-criterion space when the intercorrelation matrix is factor analyzed. After the m factors of interest have been extracted, the test constructor is required to assign a relative weight to each factor. A factor is a construct, a hypothetical entity that is assumed to underlie tests and test performance. The interpretation and naming of factors calls for psychological insights, before and after the factor analysis is made, in addition to statistical understanding. Since the test constructor is familiar with the test items, he should be able to assign meaningful names to each factor. The weighting is an indication of the relative importance of each factor to the test constructor's conceptual criterion which is an hypothetical test vector in the item and criterion space.

A rotation of the factor matrix results in factor one being collinear with the hypothetical test formed by weighting each factor. In this form, the loadings of each item on factor one is the correlation of the item with the hypothetical test. With consideration given to the size of the loading on factor one, the communality of the item and the angular displacement of the item from the hypothetical test, it is now possible to select items to construct the desired final test that will have properties similar to the hypothetical test.

Some provision must be made for up-dating the pool of items. Although items in raw score form for the same subjects can readily be added, the procedure for introducing additional items into a factor space is more complicated. Fortunately, Wherry and Winer (1953), Fruchter (1954) and Fruchter and Jennings (1962) provide partial solutions for this problem. The use of correlation matrices with missing data may be another approach to up-dating a pool of items.

CHAPTER III

MEASUREMENT IN EDUCATION AND PSYCHOLOGY

In any scientific approach, the first concern is with stating the problem in clearly defined terms. The problem should then be systematically approached within a framework of theory. A theory is defined as a deductively connected set or system of related conceptions in agreement with known properties. The body of knowledge thus acquired provides generalizations and laws that can be applied to the solution of a range of problems.

The general presentation in chapter III provides a brief outline of test theory that can be used with qualitative and quantitative criteria to develop tests. After a test has been constructed, the next logical step is to assess how well the objectives have been met that were used to prepare the test. Chapter IV contains a discussion of item and test score characteristics relevant to the assessment of tests. Validity, reliability and the related interdependencies to item and test score parameters are discussed.

A review of measurement theory and related literature is presented primarily as background material. Although the problem being investigated is related with all aspects of measurement theory, the presentation of relevant background material does not lead directly to the problem and proposed solution. However, the material presented in chapters III and IV, especially test theory, reliability and validity, is directly relevant in evaluating item selection procedures, and therefore complete

tests. Without reference to the theoretical foundation of measurement and the concepts of reliability and validity, one would find it difficult to evaluate the item selection procedures reviewed in Chapter V. Similarly, the theoretical development of the proposed selection technique and the evaluation of it follow directly from measurement theory.

A second purpose for reviewing measurement theory rests with the need to insure that users of analytic procedures for test construction are fully aware of the need to assess tests from a theoretical basis regardless of the manner by which the test was constructed. Analytic procedures must not misdirect users into believing that measurement theory is absolute or that they are not obligated to apply criteria additional to those applied analytically to the evaluation of the final test.

There is a need to continually reassess the quality of a test in terms of reliability, validity, and test score distribution whether the test items are selected by the computer or by a human. Such assessment cannot be made without a basic understanding of measurement theory.

"Measurement means the description of data in terms of numbers and this, in turn, means taking advantage of the many benefits that operations with numbers and mathematical thinking provide" (Guilford, 1954, p. 1). A product of measurement is a meaningful quantitative description given in terms that directly convey some notion of the frequency, amount or degree to which the individual manifests some property. Thus, "the scores are expressed in such a way that certain

characteristics or qualities of the individual are immediately manifest in a quantitative sense" (Ghiselli, 1964, p. 44).

In addition to quantitative description, or measurement, use is made of qualitative description, commonly referred to as classification. "All variables can be classified into one or the other of two general types, those which are qualitative variables and those which are quantitative variables" (Ghiselli, 1964, pp. 11 - 12). Qualitative variables are nominal variables whereas quantitative variables can be subdivided into ordinal variables, interval variables and ratio variables.

Measurement at best only provides information by the process of assigning numbers to individual members of a set for the purpose of indicating differences among them in the degree to which they possess the characteristic being measured. Evaluation is a judgement of merit that is sometimes based solely on measurements but more frequently involves the synthesis of various measurements and subjective impressions. Evaluation, the more recent term, includes the concept of measurement as used in education and psychology. However, measurement does not necessarily imply evaluation. "Evaluation assumes a purpose, or an idea of what is 'good' or 'desirable' from the standpoint of the individual or society or both" (Remmers and Gage, 1955, p. 21).

Test Theory

Axioms and Principal Results.

Directly related to measurement is a basic model of test theory. One fundamental notion is that any observed measurement is contaminated by an error of measurement. Thorndike (1951, p. 568) and Cronbach (1960, p. 128) have attempted to

classify these errors exhaustively. A word of caution is required here. The errors referred to are not errors due to drawing a sample from a large population of individuals. Such sampling errors are essentially independent of errors of measurement.

An extensive review and extension of classical test theory has been presented by Novick (1966). Novick attempts to show that classical test theory may be placed on a firm theoretical foundation and that its necessary assumptions are very weak and hence generally satisfied.

The simplest basic model is the classical linear model in which an observed score $\underline{X}_i$ can be divided into two additive components, a "true score" $\underline{T}_i$ and an "error score" $\underline{E}_i$, that is

$$\underline{X}_i = T_i + E_i$$

It is assumed, for $(\underline{E}_i)$, that we are dealing with random errors, normally distributed, where (a) the mean $\underline{E}(\underline{E}) = 0$, (b) covariance $\underline{E}(\underline{T}_i, \underline{E}_i) = 0$, (c) covariance $\underline{E}(\underline{E}_1, \underline{E}_2) = 0$. $\underline{E}_1$ and $\underline{E}_2$ are random errors on two testing occasions (Gulliksen, 1950 a). $\underline{E}$ denotes expected values. The variance of the gross observed scores is then given by

$$s_x^2 = s_t^2 + s_e^2$$

Gulliksen (1950 a) has shown that the index of reliability for a test is the proportion of true score variance divided by the observed score variance, that is

$$r_{x_g x_h} = \frac{s_t^2}{s_x^2}$$

where $\underline{x}_g$ and $\underline{x}_h$ are two parallel form measures. It may be shown that

$$s_x^2 = s_x^2 r_{x_g x_h} + s_e^2 ,$$

and that

$$s_e = s_x \sqrt{1 - r_{x_g x_h}} ,$$

where $\underline{s}_e$ is the standard error of measurement. This is a fundamental concept in test theory and defines an important characteristic of a test.

However, validity is the most important criterion by which a test may be judged (Helmstadter, 1964). Validity can be regarded as being composed of essentially two components: the accuracy of measurement or reliability and what the test intended to measure or the criterion for the relevance of the test (Cureton, 1950; Remmers and Gage, 1955).

Test Score. "A score is a number assigned to an examinee to provide a quantitative description of his performance on a particular test" (Ebel, 1965, p. 462). When a test contains many items, the raw score of an individual is commonly defined as the number of items that are answered correctly. A correction for guessing or a differential weighting system for the items may be applied to improve or refine the raw score. There is, however, not complete agreement among test specialists concerning the questions of using a weighting technique or a correction for guessing (Traxler, 1951).

Test Score Distribution. Distributions of scores vary markedly in their shape, manifesting different degrees and combinations of skew-

ness and kurtosis. Early investigators seemed to think that there was a natural law for human abilities to be normally distributed. Now, it is realized that such a statement is meaningless since the shape of a distribution depends on the scale of measurement.

Moment statistics can be used to summarize and characterize data. The most important set of moments in statistical theory is obtained by calculating moments about the arithmetic mean. Two of them, the arithmetic mean and the variance are in common use. The first four moments, commonly called deviations from the mean or simply deviations, are defined as follows:

$$\mu_1 = \frac{\Sigma x}{N} = 0$$

$$\mu_2 = \frac{\Sigma x^2}{N} = s^2$$

$$\mu_3 = \frac{\Sigma x^3}{N}$$

$$\mu_4 = \frac{\Sigma x^4}{N}$$

where $\underline{x}$ represents the deviation of each score from the mean of all the scores.

A measure of skewness defined in terms of moments is

$$g_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

The value of g_1 will be zero for symmetrical distributions. Skewness, measured as a departure of g_1 from zero, is positive when g_1 is positive and negative when g_1 is negative.

The degree of kurtosis can be described by

$$g_2 = \frac{\mu_4}{\mu_2^2} - 3$$

A distribution of scores is leptokurtic when g_2 is positive, platykurtic when g_2 is negative, and normal when $g_2 = 0$.

A normal frequency distribution can be completely described by the mean and the variance when both g_1 and g_2 are zero. While it is convenient to use the normal curve, one must remember that "very few of the instruments used in psychological 'measurement' involve equal unit scales - the measuring units are frequently arbitrary or even accidental" (McNemar, 1962, p. 28). It would seem that skewness and kurtosis are partly a function of the accidental nature of the measuring units. The values of g_1 and g_2 are, however, useful for descriptive purposes.

The higher moments

> . . . have relatively little use in elementary applications of statistics, but they are important for mathematical statisticians in the study of the properties of distributions and in arriving at theoretical distributions fitting observed data (Hays, 1963, p. 186).

The means, standard deviations and intercorrelations of items in a test have a very important bearing upon the shape of the total-score distribution. If the items are relatively easy, a negatively skewed distribution will result, whereas, if the average item mean (item difficulty) becomes lower the score distribution becomes positively skewed. With items of medium difficulty, the distribution becomes symmetrical.

The chief effect of item intercorrelations is upon kurtosis. As item intercorrelations increase, the distribution of total-scores

grows flatter from mesokurtic to platylkurtic, to rectangular, to bimodal and finally U-shaped (Guilford, 1954). When item intercorrelations increase, the test reliability subsequently increases which usually influences the validity coefficient.

Thus, the distributions of actual test scores depend upon the way the test is constructed. Although

> relatively little of a precise nature is now known regarding the effect of item selection on test skewness, kurtosis, or on the constancy of the error of measurement throughout the test score range, . . . it is possible, however, to select items in such a way as to influence the test mean, variance, reliability and validity (Gulliksen, 1950 a, p. 365).

This in turn will directly influence the test score distribution. Mollenkopf (1949, 1950) has shown that the variation of the error of measurement with test score depends on the third and fourth moments. This offers some difficulties in the theoretical analysis of item selection procedures. As a partial aid to the solution of the above problem, Ray, Hundleby and Goldstein (1962) demonstrated that indices of skewness and kurtosis for a test score distribution can be expressed in terms of item parameters.

Although attempts have been made to select items on the basis of the first four moments, the selection of items to form a test with given skewness and kurtosis has not been solved. Ray, Hundleby and Goldstein (1962) claim that any moment employed in describing the frequency distribution of raw scores can be expressed as a function of item parameters but they do not show how this information can be used in the practical case of selecting items from a predefined pool to construct a test. Since the correlation between gross scores is identical with

the correlation between linear transformations of gross scores, the equations dealing with the effect of the test length and group heterogeneity on reliability and validity hold for gross scores and for any linear transformation of gross scores.

The shape of the score distribution may be altered by using various transformations. One of the most frequently used is a logarithmetic transformation of a psychological variable to obtain scores that are at least approximately normal (McNemar, 1962). Use of the normal curve is merely a convenience and is not necessarily based on any "normal distribution of behaviour" in nature. Since the normal frequency distribution has commonly been found to be characteristic, or nearly so, of the distributions of scores on a wide variety of characteristics, it has been established as one particular distribution to be used as a frame of reference for comparison purposes. The normal frequency distribution has also been termed the curve of error (Ghiselli, 1964, p. 59) since it is closely approximated in situations where a score is determined by a large number of factors which operate under conditions of equal likelihood.

Ghiselli draws the conclusion that

> . . . there are, of course, a wide variety of differently shaped distributions that could be adopted as the theoretical model of the distribution of psychological traits. Of all the possible distributions there appears to be more basis for choosing the normal frequency distribution (Ghiselli, 1964, p. 62).

Standard Error of Measurement. The standard error of measurement is an estimate of the standard deviation of the errors of measurement associated with the test scores in a given set. In terms of the

reliability coefficient, r_{xx}, and the standard deviation, s_x, the standard error of measurement formula presented by McNemar (1962) is

$$s_e = s_x \sqrt{1 - r_{xx}}$$

Thus, s_e is useful in establishing 'true score' limits.

Since the reliability coefficient is dependent on the variability of the group to which the test is applied, whereas the standard error of measurement is affected very little by this characteristic, the latter is sometimes proposed as a measure of reliability (Ebel, 1965). However, use of the standard error of measurement often assumes that the error in estimating the true score is the same in all parts of the range of the observed score. This by no means is necessarily true. Also, for tests using a given type of item, the standard error of measurement is almost entirely dependent upon the the number of items in the test and minimally upon their quality (Lord, 1957; Lord, 1959; Swineford, 1959).

With zero skewness and kurtosis of 3, the error of measurement is constant with respect to size of test score (Gulliksen, 1950 a). Mollenkopf (1949, 1950) has provided empirical evidence to show that the error of measurement is affected by the effects of skewness and variations in kurtosis. He concluded that slight skewing could be tolerated but not departures in kurtosis from 3 as the error of measurement will then vary with the magnitude of the test score. Lord (1952) has suggested that the dispersion of errors will be smallest at the tails of a distribution and that the standard error of measurement should be considered as an average error.

Transformations of Scores. Since many raw score measurements do not have the characteristics of a desirable system of units, raw scores are often changed by means of a transformation to "transmuted scores". This may permit easier interpretation of the score and allow comparisons to be made between different tests or between different parts of the same test. A distribution of raw scores is frequently converted to a set of norms since "a raw score on any psychological test is, in itself, quite meaningless" (Anastasi, 1961, p. 76). There are various ways in which raw scores may be converted. DuBois (1965) defines two general classes of norms; reference norms and statistical norms.

Reference norms are those which have raw scores translated into meaningful work standards closely related to psychological tests. These include work norms, grade norms, mental age norms (MA) and chronological age norms (CA). Work norms are expressed in units of production in a standard time interval by a member of a specified group. In age norms, the mean performance for each age is calculated and subsequently used to construct a distribution of scores from which to estimate an age equivalent. Quotient norms have been common in mental testing such as for example the intelligence quotient. The trend in mental testing now seems to be towards the use of statistical norms, rather than reference norms. Wechsler (1958) and Terman and Merrill (1960) have provided statistical norms for the Stanford-Binet Intelligence Test, the Wechsler Intelligence Scale for Children and the Wechsler Adult Intelligence Scale.

When mathematical transformations are applied to raw scores in calculating statistical norms, the norms are useful for comparison purposes but have in and of themselves no direct meaning. The three main types of statistical norms are percentiles, standard scores and normalized scores which differ primarily in the shape of their distributions. The distribution of percentiles is theoretically rectangular where 1 percent of the sample size is included between two adjacent percentiles. The shape of a distribution of standard scores is identical to the distribution of raw scores. In general, if we wish to transform a set of scores, $\underline{X}$, having a mean, $\underline{M}_x$, and a standard deviation, $\underline{s}_x$, to new values, $\underline{Y}$, with mean equal to any value, $\underline{M}_y$, and a standard deviation, $\underline{s}_y$, we can apply the formula

$$Y = \frac{s_y}{s_x} X - \frac{M_x}{s_x} s_y + M_y$$

Three common sets of standard scores are (a) standard z scores (0, 1), (b) T scores (50, 10) and (c) stanines (5, 1.96). Normalized scores are similar to standard scores with respect to characteristics of the mean and standard deviation. An additional property of correction for departures from normality are made on the original raw scores. The distribution of the normalized scores approximates the normal distribution with decreasing "goodness of fit" as the shape of the original distribution departs from normality.

Various types of samples, such as male or female college students, are used as a basis for establishing norms. Adequate norms for a special selected group may be calculated by using a large number of cases and a

representative sample.

Test Development

"A test is a general term used to designate any kind of device or procedure for measuring ability, achievement, interest and other traits" (Ebel, 1965, p. 466). The construction of any test involves numerous decisions. In the preparation of a test, one of the most important yet most often neglected aspects has been a careful delimitation and breakdown of the area or trait involved (Helmstadter, 1964). A test should be based on a representative sampling of the content studied while having a representative sampling of the abilities or skills emphasized in the course (Adams and Torgerson, 1964, p. 322). As no single instrument can measure all skills over an entire content area, resort must be made to the procedure of using a representative sample of test items. Ultimately, the test constructor in applying his experience and judgemental skill, decides exactly what will or will not be included in the measure. What constitutes important materials can only be determined by careful attention to the goals of a course. Part of this decision should be determined by reference to future courses or types of employment that the examinees will enter.

Qualitative Criteria. The plan for a test should consider the relative emphasis to be given both to content areas and to the processes or cognitive abilities which are specific ways of responding to or dealing with course content. A detailed analysis of educational objectives for student achievement has been edited by Bloom (1956). Bloom and associates have developed a taxonomy of educational objectives under which

educational goals and test items in the cognitive areas may be classified. The major categories of the Taxonomy, in increasing degrees of complexity are (a) knowledge, (b) comprehension, (c) application, (d) analysis (e) synthesis, (f) evaluation.

Stoker and Kropp (1964) report general support for the hierarchical structure of the cognitive process if evaluation is placed before synthesis. Additional support for Bloom's notion of hierarchical structure is provided by Ayers (1966). The results from a factor analytic study by Ayers are in general agreement with a hierarchical nature but there is some question as to whether or not the same factors and hierarchical order, as that presented by Bloom, will be confirmed.

Suggestions for preparing good test items can be found in several books (Lindquist, 1951; Thorndike and Hagen, 1955; Ebel, 1965). A list of additional references for item writing in various subject fields has been prepared by Adams and Torgerson (1964, pp. 396 - 399).

Test items have frequently been dichotomized into essay test items and objective test items. In this setting, essay is intended to include other supply-type test items such as completion questions. Objective items, which can be thought of as choice-type instead of supply-type, can be subdivided into true-false items, multiple-choice items and matching exercises. "There is a growing recognition that many of the criticisms of both approaches are not necessarily inherent but grow out of ineffectiveness in their application" (Adams and Torgerson, 1964, p. 332).

In the essay test, a few questions or problems are presented and students are asked to supply the answers. A large number of questions, with a limited number of alternative answers for each, are used in objective

tests. It is comparatively easy to construct an essay test but difficult to grade for more than a few students. The multiple-choice test is relatively more difficult to construct but can be graded easily for many students. An essay test is usually less reliable than a multiple-choice test because of the minimal sampling of content and variability in scoring of questions. Although well-constructed multiple-choice tests are accepted as effective measurement instruments (Ebel, 1965), they are often criticized as measuring only the simple facts of subject matter and thus provide no evidence regarding command of cognitive abilities of greater complexity. Also, multiple-choice tests are regarded by critics (Hoffman, 1962) as being only a measure of memory rather than understanding. Essay tests can, it is maintained, be used to allow a student to demonstrate his ability to organize and present a creative answer. Rather than try to decide whether multiple-choice examinations are generally better than tests of the essay type, or vice versa, it would be more appropriate to see how they both can be made as effective as possible and how they can be used to complement one another. Ebel (1965, pp. 109 - 110) has outlined how essay and objective tests are useful for different purposes and in different situations.

The use of multiple-choice items, where an item is scored either 1 or 0, introduces the problem of what level and distribution of difficulties are appropriate for the questions included in the test. One answer is to include only questions that most students should, in the teacher's opinion, be able to answer. If this is done, many students will answer most questions correctly resulting in poor discrimination among students on level of achievement. Another alternative is to use

items on which approximately half the students are successful. This approach will contribute the most information as to relative levels of achievement among the students tested. When the difficulty level of an item, p, is .5, the maximum possible item variance, s, is obtained by $s = \sqrt{p\,q}$, where $q = 1 - p$. Departures from $p = .5$ will result in a decreased item variance. Although departures from $p = .5$ may yield more reliable scores for the same amount of testing time, an optimal psychometric situation where $p = .5$ may prove to be more worrisome to the students. When $p = .5$, half of the students will fail any item resulting in a mean score of only 50 percent. It should be noted that p, an average item score, is also an average index of item difficulty for individuals. Coombs (1950) has commented on the fact that the difficulty of an item varies for different individuals. The index p does not yield accurate information concerning the item's difficulty for a given individual.

"There is no formula for determining the exact distribution of item difficulties" (Freeman, 1955, p. 39). The determination of the optimum difficulty of the test items to be used in a test is a problem on which there is not complete agreement among test specialists. Some test authorities prefer approximately equal numbers of items at all levels arranged from very easy to very difficult (Remmers and Gage, 1955; Nunnally, 1959), others prefer to have the majority of items near the 50 percent difficulty level. Richardson, for example, found that

> . . . a test composed of items of 50 percent difficulty has a general validity which is higher than tests composed of items of any other degree of difficulty (Richardson, 1936, p. 47).

Gulliksen, in a theoretical analysis, concluded that

> In order to maximize the reliability and variance of a test the items should have high intercorrelations, all items should be of the same difficulty level, and the level should be as near 50 percent as possible (Gulliksen, 1945, p. 79).

In spite of the fact that the maximum item criterion correlation occurs for items of 50 percent difficulty, another special level of difficulty may prove to be valuable in a particular situation. When items have low intercorrelations, a distribution of item difficulties clustered around the 50 percent level often approximates the distribution required to obtain maximum discrimination throughout the range of scores. The distribution of difficulty indices should be made more platykurtic or rectangular than usual if equal accuracy of measurement and discrimination are desired throughout the range of scores for items with relatively high intercorrelations. An extended discussion of the above statements has been prepared by Brogden (1946). When selecting a specific group of subjects, Lord (1953) suggests that the average item difficulty should match the selection ratio. If the top 30 percent of persons were to be selected, the most efficient test would be that for which the average item difficulty is at 30 percent.

The general procedure in common practice "in the arrangement of standardized test items tends to follow the procedure of presenting items covering a wide range of difficulties in ascending order from the very easy to the most difficult" (Greene, Jorgensen and Gerberich, 1954, p. 91).

Apart from statistical decisions the onus is on the test constructor to select the desired level or levels of item difficulty, to

suggest whether or not a weighting of items is necessary, to decide whether or not to correct for guessing, to accept or reject a given level of reliability and/or validity coefficient, and to make decisions on a host of other major considerations in preparing a test. No present statistical technique can replace the judgement of the subject matter expert in the selection and rejection of items to sample representative content domains and educational objectives.

Quantitative Criteria. In developing a test consideration is given to the statistical properties obtained through item analysis for each individual item obtained by an item analysis. When the decisions have been made by "experts" on cut-off points, it is relatively easy to select items with the desired properties. The items selected, on the basis of item characteristics, can now be used to form an item pool. It would be desirable to select a sample of items from the item pool that would result in a desired mean, variance, skewness, kurtosis and distribution shape. This is, as yet, not possible. The most frequently used procedure at present for constructing a test is based upon the results obtained from an item analysis.

In many situations item pools are constructed by selecting items on the basis of an item analysis of several tests administered to different groups of subjects. Tests are subsequently constructed by selecting items on the basis of the initial item analysis used to construct the item pool. However, it must be noted that item analysis results for an item are always specific to the particular group and the particular subset of items into which the item is embedded. Thus,

the item analysis for a sample of items from the item pool might well differ from those used to construct the pool of items. On this basis therefore, it seems reasonable that whenever possible the initial item pool should be re-evaluated in terms of the user's needs.

CHAPTER IV

ASSESSMENT OF TESTS

Whether a test is "hand made" or developed according to some analytical criterion, all tests must be subjected to the same criterion for their evaluation. A multitude of approaches are available for assessment of tests.

The most recent reference that provides a general consensus by authorities in the field of measurement regarding what and how to evaluate tests is Standards for Educational and Psychological Tests and Manuals (Standards) (1966). Although the presentation in the Standards is very brief and must therefore be supplemented with material from other publications, it should be considered as an authoritative voice in deciding what is relevant or nonrelevant in evaluating a test. As an aid to test development, the Standards provide a kind of check-list of factors to be considered in designing the standardization and validation of tests. The main topics covered are: (a) dissemination of information, (b) interpretation, (c) validity, (d) reliability, (e) administration and scoring, and (f) scales and norms.

Validity

Test validity is concerned with what a test measures and how well it does so. Validity is a complex concept that has been interpreted in various ways by different writers. Many types of validity and their general classifications have been described. Thorndike and Hagen (1955) suggest a dichotomy of types of validity: validity which

depends primarily upon rational analysis and professional judgement, and that which depends upon empirical and statistical evidence. The dichotomy, similar to that above, proposed by Ebel (1965) is, respectively, concerned with primary or direct validity as contrasted with secondary or derived validity.

Some types of validity, reviewed by Ebel (1965), that seem appropriate for each category are listed below:

Direct	Derived
Validity by definition	Empirical Validity
Content Validity	Concurrent Validity
Curricular Validity	Predictive Validity
Intrinsic Validity	Factorial Validity
Face Validity	Construct Validity

The distinction between the two categories is not explicit nor clearly defined since factorial validity and construct validity, "despite their involvement of multiple measurements and coefficients of correlation, do represent a basic (primary) kind of validity" (Ebel, 1965, pp. 381-382). A standard reference, Technical Recommendations for Psychological Tests and Diagnostic Techniques (1954), has the various types of validity classified under four categories, designated as content, predictive, concurrent, and construct validity. These four aspects of validity have been used as a basis for developing more elaborate sub-classifications.

In Standards for Educational and Psychological Tests and Manuals (1966), a revision of two documents: (a) Technical Recommendations for

Psychological Tests and Diagnostic Techniques (1954) and (b) Technical Recommendations for Achievement Tests (1955), three kinds of validity coefficients are distinguished. The three aspects of validity corresponding to three aims of testing may be designated as follows:

1. Content Validity - The test user wishes to determine how an individual performs at present in a universe of situations that the test situation is claimed to represent.

2. Criterion-Related Validity - The test user wishes to forecast an individual's present standing on some variable of particular significance that is different from the test.

3. Construct Validity - The test user wishes to infer the degree to which the individual possesses some hypothetical trait or quality (construct) presumed to be reflected in the test performance. (Standards for Educational and Psychological Tests and Manuals, 1966, p. 12)

"Probably the most sophisticated form of content validity is that which makes use of the technique called factor analysis" (Helmstadter, 1964, p. 92). In like manner, Guilford maintains that "the best answer to the question, "What does this test measure?" is in the form of a list of primary factors with which it correlates and their proportions of variance in the test" (Guilford, 1965, p. 472). The above validity estimate is known as factorial validity. According to Guilford (1965) this type of validity is basic to the understanding of other kinds of validity and of many phenomena of correlation in general.

Whereas predictive and concurrent validation are judged for a test by a statistical study of results, content validity is established by logical examination of the test and the methods used in its

preparation. Subjective judgement, "be it termed professional judgement, common sense, or 'expertese', is involved in all phases of content validity and is its paramount characteristic" (Ghiselli, 1964, p. 345).

Although subjective judgement and factorial validity seem to represent, respectively, an evaluative position based on personal opinion versus an objective statistical solution, subjective judgement plays a prominent role in factorial validity. The postulated constructs represented by each of the factors resulting from a factor analysis are defined, in the main, by persons familiar with the variables used in the particular analysis being considered.

Emphasis has been given to content validity because of its basic position in all measurement problems. Since test questions are only a sample of all possible questions that might be asked, items may or may not be representative of the total domain of appropriate questions. In an ideal situation, a test constructor should define a subset of the universe to be studied, e.g., an outline of the course content should be used in preparing an achievement test, from which a sample of items is selected to represent the content. Test developers should exercise great care to match their achievement tests to the course of study. Item sampling is sometimes very poor in tests constructed by an inexperienced or untrained tester.

Content validity requires judging whether each item, and the distribution of items as a whole, covers the subject matter of interest to the tester. The decision to accept or reject an item, on the basis of its content, remains with the test user rather than the test

constructor. Although the test constructor can state the source of his items, they will rarely correspond perfectly to what the tester requires. Thus, it would appear that content validity is one type of validity with which we should be deeply concerned. The assumptions underlying the use of content validity have been summarized by Lennon (1956).

Two approaches are used in calculating a criterion-related validity coefficient. The procedure is essentially a measure of statistical relationship between test scores and one or more external variables considered to provide a direct measure of the characteristic or behaviour being evaluated. If a test is to be used for assessment of present status, the criterion data should be collected concurrently with the testing. For predictive purposes, the criterion data would usually be collected at a later time.

Cronbach and Meehl (1955) presented the notion of construct validity which has been formally adopted by the American Psychological Association, the American Educational Research Association and the National Council on Measurement in Education[1]. A combination of logical and empirical attack is required in gathering data to examine construct validity. Although construct validity, as a concept, appears to be fully acceptable to many authoritative psychometricians, Horst maintains that "it is very difficult to incorporate it (construct validity) in or integrate it with a logical and practical theory of measurement" (1966, p. 346). While there may be problems associated with using the concept of construct validity in measurement theory, the

---

[1]See Standards for Educational and Psychological Tests and Manuals (1966).

general consensus appears to be that of retaining the term and the theoretical framework upon which the notion rests.

The emphasis in the definition of validity is upon what is being measured. It must be emphasized that there is no one measure of validity. A test or scale is valid for the particular scientific or practical purpose of its user. Thus, different types of investigation are required to establish the validities when several types of criteria are involved. The procedure for establishing criterion-related validity differs from the approach used to determine construct validity which in turn differs from how content validity is established for a test. When assessing the validity of a test, the question "Valid for what?" should be answered.

Reliability

"The reliability of any set of measures is logically defined as the proportion of their variance that is true variance" (Guilford, 1965, p. 439), whereas the index of reliability is the "correlation between true and observed scores" (Gulliksen, 1950 a, p. 22). When reliability is defined as the ratio of the true score variance to observed-score variance in the population, the ratio is sometimes known as an intra-class correlation.

Traditionally, reliability, a generic term referring to many types of evidence, is concerned with the question "How consistently does a test measure?" Several approaches to score consistency results in several types of reliability coefficients. All types do not answer the same questions. As a result of inconsistency in terminology used

by researchers and vague definitions of terms, a joint committee of the American Psychological Association, American Educational Research Association, and National Council of Measurements Used in Education prepared a publication entitled Technical Recommendations for Psychological Tests and Diagnostic Techniques (1954). An attempt was made to standardize and classify the various types of reliability. The three main subclassifications are as follows: (a) A measure based on internal analysis of data obtained on a single trial of a test is to be known as a coefficient of internal consistency. The most prominent of these are the analysis of variance method (Kuder and Richardson (1937) and Hoyt (1941)) and the split half method, (b) a coefficient of equivalence is obtained by calculating a correlation between scores from two forms given at essentially the same time, and (c) the correlation between test and retest, with an intervening period of time, is a coefficient of stability. The latter procedure may be used with parallel forms of a test or a second administration of the same test after an intervening period of time.

Cronbach (1951), using one of the reliability formulas that was derived from a more general theoretical approach by Kuder and Richardson (1937), designated a particular reliability index as "coefficient α" which would replace the name, "Kuder-Richardson formula number 20", now commonly used. Attractive features of the formula used to calculate coefficient α are that it yields the mean of the correlations resulting from all possible ways of splitting a given test into two halves and that it gives the proportion of first-factor variance extracted from the inter-correlations of the test items. An additional feature of the formula used to calculate coefficient α is that the formula is not restricted to

items scored 0 and 1.

The reliability of a test is often referred to as being a measure of internal consistency, rather than a temporal (retest) index, which seems to follow logically from classical test score theory (Baggaley, 1964). This is further reflected in the observation that "in developing the vast majority of tests constructed today the makers strive towards internal consistency" (Guilford, 1954, p. 388).

However, while Guilford (1954) maintains that reliability is the minimum information one should have concerning a test, he further suggests that it is certainly not the most useful information. "It is sometimes said that reliability is important because it contributes to validity and that validity is the important goal" (Guilford, 1954, p. 389). Thus, "a test cannot measure more accurately what it is intended to measure than the accuracy with which it measures what it does measure. Hence in order to be valid a test must be reliable" (Ebel, 1965, p. 389). Therefore, to be concerned with test validity directly implies a consideration for the reliability of a test. "Reliability is a necessary condition for validity in an educational achievement test, but it is not a sufficient condition" (Ebel, 1965, p. 309).

In the present set of Standards, the reliability coefficients are not classified into several types as in the Technical Recommendations. The explanation given for this move is that the "terminological system breaks down as more adequate statistical analyses are applied and methods are more adequately described" (Standards for Educational and Psychological Tests and Manuals, 1966, p. 26). It is recommended in the Standards that test authors work out suitable phrases to convey the

meaning of whatever coefficient they report. The rationale for the presentation of a descriptive rather than a categorized type of reliability coefficient is that different methods take account of different sources of error, which when clearly labeled, is the most informative outcome of a reliability study. If this approach is used, it is imperative that the method used to derive the reliability coefficient be clearly described. The impetus for this trend appears to have resulted from suggestions made by Cronbach et. al. (1963) in Theory of Generalizability: A Liberation of Reliability Theory.

Interdependencies

Factors Affecting Reliability and Validity. Reliability is dependent upon various determining factors, such as speed of work, heterogeneity of subjects, length of test, difficulty level of the items, and approach used to estimate reliability. In general, reliability is a function of item by person tested. The parallel form estimate of reliability is often considered to be a lower bound because it includes form to form and time fluctuations in its definition of error. For the above reasons, a parallel form estimate is often the preferred measure (Helmstadter, 1964). Split-half reliability is usually regarded as representing the upper bound of the true reliability. This is especially relevant when applied to tests having a large speed component. Homogeneous tests are likely to be more reliable than heterogeneous tests whereas scores obtained from heterogeneous groups are likely to be more reliable than scores obtained from homogeneous groups (Ebel, 1965). As the length of a test is increased, the reliability of the test increases. The relationship between test reliability and

test length is expressed by the generalized Spearman-Brown formula (Gulliksen, 1950 a). "Contrary to popular belief, a good test seldom needs to include items which vary in difficulty" (Ebel, 1965, p. 339). When items have a difficulty level of .5, more variable scores are obtained from a test. The reliability of a test is likely to be higher when there is a maximum score variance resulting from the use of items having difficulty indices near .5.

The Kuder-Richardson Formulas for estimating the reliability of a test, r_{xx}, depend upon item statistics. They were developed because of dis-satisfaction with split-half methods. The use of item statistics removes such biases as may arise from arbitrary splitting into halves. When an accurate and practical formula is required, calculation of the reliability coefficient for a test is generally estimated by using the Kuder-Richardson 20 (KR-20) formula. The relationship of item analysis data to reliability may, perhaps, most clearly be demonstrated by means of the following equation. The expression is

$$r_{xx} = \frac{K}{K-1}\left[1 - \frac{\sum_{g=1}^{k} s_g^2}{\left(\sum_{g=1}^{k} r_{xg} s_g\right)^2}\right]$$

where K is the number of items in the test,

S_g^2 is the item variance which equals $p_g - p_g^2$,

P_g is the difficulty of item g,

$r_{xg}\ S_g$ is the item reliability index, and

r_{xx} is the reliability of the total test.

Item "g" is an item in test "x". Although the KR-20 formula yields accurate results, considerable work is required in calculating r_{xx}. The most common modified KR formula proposed is that known as the KR-21 formula. If the item difficulties are very nearly equal, the KR-21 formula will provide a quick estimate of the lower bound of r_{xx}. This formula only requires information regarding the test mean, variance of the raw scores and the number of items in the test. The estimate obtained by using KR-21 is generally lower than that calculated by using formula KR-20 whereas the odd-even estimate will generally be higher than the KR-20 value.

The corresponding general formula for validity is presented below. We have

$$r_{xy} = \frac{\sum_{g=1}^{K} r_{yg} S_g}{\sum_{g=1}^{K} r_{xg} S_g}$$

where r_{yg} is the point biserial correlation of item g with the criterion y,

r_{xg} is the point biserial correlation of item g with the test x, and

r_{xy} is the correlation between the criterion and test (Gulliksen, 1950 a).

Since transformations of test scores can be used to obtain scores with a specified mean, variance and, within certain limits, the form of score distribution, it is suggested that test construction procedures may profit when an emphasis is placed upon producing reliable and valid tests. An attempt to produce a single test that is

both highly reliable (in the internal consistency sense) and also highly valid is truly a meritorious task. Unfortunately the two goals are incompatable in some respects. The requirements, as outlined by Guilford, for maximal reliability and predictive validity are as follows:

> Maximal reliability (internal consistency type) requires high intercorrelation among items; maximal predictive validity requires low intercorrelations. Maximal reliability requires items of equal difficulty; maximal predictive validity requires items differing in difficulty (Guilford, 1965, p. 481)

Thus, there must be some compromising of aims since both reliability and validity cannot be maximal especially when there is a restriction on the number of items used to construct a test. An optimal situation may be to treat both properties with equal emphasis. However, to "err on the side of (high) validity, which after all, is the more important" (Guilford, 1965, p. 481) will probably lead to the construction of a highly acceptable test.

The number of measurements (items) used in constructing a test will influence results calculated from the data. Gulliksen has shown that "increasing the length of a test $\underline{K}$ times multiplies the mean by $\underline{K}$, provided that each of the new parts is parallel to the original" (Gulliksen, 1950 a, p. 69). Lengthening a test $\underline{K}$ times increases the variance of gross scores as indicated in the following equation

$$S_c^2 = S_1^2 K [1 + (K - 1) r_{12}]$$

where

$\underline{S_1^2}$ is the variance of the unit length test,

$\underline{K}$ is the ratio of the number of items in the new test to the

number in the unit length test,

$\underline{r}_{12}$ is the correlation between the two parts, and

S_c^2 is the variance of the lengthened test.

Similarly,

$$S_c = S_1 \sqrt{K + K (K - 1) r_{11}}$$

is the formula relating the increased length of a test to its standard deviation ($\underline{S}_c$) where $\underline{r}_{11}$ is the reliability of the unit test (Gulliksen, 1950 a).

The effect of test length on reliability is given by

$$R_{kk} = \frac{K r_{11}}{1 + (K - 1) r_{11}}$$

which is known as the general Spearman-Brown formula where $\underline{R}_{kk}$ is the reliability of the lengthened test. The relationship between test length and its validity is given by

$$R_{KI} = \frac{r_{1I} \sqrt{K}}{\sqrt{1 + (K - 1) r_{11}}}$$

where $\underline{r}_{1I}$ is the validity coefficient of the unit test and $\underline{R}_{KI}$ is the augmented validity coefficient. As the test length is increased, reliabilities approach unity. However, in contrast to reliability "the validity coefficient is usually considerably smaller than the test reliability (which) usually means that changing the length of a test can be expected to have only a very slight effect on the validity of the test" (Gulliksen, 1950 a, p. 90).

In a discussion of reliability and validity, Rozeboom (1966, p. 422) says that "it is debatable whether the practical benefits of reliability theory are sufficiently bountiful to recompense the labour that test theorists have invested in it. The primary justification of reliability theory lies in abstract curiosity". The one useful aspect of the test's reliability index is that it is an upper limit to its validity. Thus, information about test reliability is especially important when data are not available on a test's validity for its intended purpose. When validity estimates are available, the test's reliability is a matter of indifference (Rozeboom, 1966).

The problem of obtaining a suitable criterion arises whenever a prediction is to be made. At times samples are selected that have a marked restriction of range on the resulting test scores. A failure to cross-validate can lead to exaggerated claims as to the effectiveness of the prediction or selection. Apart from the criterion problem, there are the issues of guessing and faking, and response sets.

The Criterion. "The so-called criterion problem refers to the fact that in many cases it is extremely difficult to obtain adequate evidence for the validity of a test because no criterion appears to be completely satisfactory" (Helmstadter, 1964, p. 145). Since the concept of predictive validity involves the correlation of a psychological measure with a special kind of measure called the criterion, one of the first tasks is to define a conceptual criterion by means of verbal statements from which a criterion measure is developed that is stated in operational terms. "The only method for "validating" a criterion measure

is a logical analysis of its relevance to the conceptual criterion" (Astin, 1964, p. 811).

An outline of the problems involved in the use of criterion measures has been subdivided into three general categories as follows:

1. The Nature and Role of the Criterion (definitions, common fallacies about criteria, and certain logical and technical considerations in developing criterion measures).

2. Criteria and Test Development (the function of criteria in the construction and validation of tests).

3. Criterion-Centered Research versus Construct Validity (similarities and differences between the two approaches, and the case for criterion-centered research). (Astin, 1964, p. 807)

Cureton (1951, pp. 626 - 674) and Horst (1966, pp. 334 - 347) have directly related the criterion problem to test validity in detailed presentations both entitled Validity.

When a test is being constructed, the ends desired in an applied setting should first be established by defining those ends in terms of a set of criteria. Thus, specification of conceptual criteria and some attempt at criterion development appear to be important preliminaries to the construction of any test which is designated for applied use.

Item Analysis. "The Major goals of item analysis are the improvement of total-score reliability or of total-score validity, or both, and the achievement of better item sequences and types of score distributions" (Guilford, 1965, p. 493). The commonly used descriptive statistics for item parameters are:

1. The proportion of persons answering each item correctly. This quantity is a measure of item difficulty.

2. The reliability index, which is the point-biserial correlation between item and total score multipled by the item standard deviation. A reliability index is not equivalent to the index of reliability.
3. The validity index, which is the point-biserial correlation between item and criterion score multipled by the item standard deviation. (Gulliksen, 1950 a, p. 385)

An item analysis essentially provides two kinds of information. It provides an index of item difficulty and an index of validity, where the term validity is used in a very broad sense. These indices may show how well the item discriminates in agreement with the rest of the test, generally the total test score as an internal criterion, or how well it predicts some external criterion. Item validity is thus a case of construct validity when the criterion is the total score and predictive validity when one uses an external criterion. The homogeneity (internal consistency) of a test is increased when items are selected which correlate highly with total score.

Short-cut methods of estimating these parameters from a portion of the data have been presented by several writers. Kelly (1939) suggested that two special criterion groups be formed: an upper group, consisting of 27 percent of the total group, who received the highest total test scores and a lower group consisting of an equal number from those who received lowest scores. Item analysis data would be calculated from this portion of the data. Graphic procedures may be used to calculate item difficulty and item discrimination indices. Guilford (1954)

and Helmstadter (1964) provide a detailed explanation of these techniques.

However, with the increasing use of modern electronic machines, short-cut methods will become increasingly less desirable since the computational labour is reduced considerably and more information is provided by using all responses in a more rigorous procedure.

The difficulty level of items will determine, to a large extent, the shape of the test score distribution. If a multiple-choice test is used, the number of alternatives used for each item usually restricts the range of probable scores. The range would be from approximately 20 percent correct answers to 100 percent correct answers, allowing for random guessing, when using five alternatives. In a practical test situation this would undoubtedly vary since the alternative selected is not generally made in a purely random fashion. One method of selecting items to result in a wide spread of test scores is to use the rule of selecting a set of items whose average difficulty level is near the middle of the possible score range. Although "the ideal distribution of difficulties varies in terms of the use which will be made of the test and the intercorrelations of the items . . . a good general procedure is to choose approximately an equal number of items at each difficulty level in the possible score range" (Nunnally, 1959, pp. 146 - 147).

While the multiple correlation approach is generally accepted as being the superior method for predicting a criterion from a test composed of several items, in order to maximize validity, the procedure most frequently used to select items to form a test is that based on item analysis. After the item characteristics have been determined,

either analytically or graphically, from the empirical tryout of the pool, a number of statistical procedures can be used to help construct a test. "The purpose of item analysis is to select from an item pool a minimum number of items which will give a maximum prediction of a criterion" (Nunnally, 1959, p. 144).

A guide for selecting items to construct a test is to use, in general, items in the difficulty range of .20 to .80. Items more difficult than the .20 level would not likely be answered by many students whereas items of .80 difficulty or greater may be so easy that one is only adding a constant to the individual's score since nearly everyone receives credit for this question. Several relationships for a discrimination index have been presented. Various combinations of proportions calculated for the upper and the lower groups have been used. A correlation between criterion or total score and item is sometimes used. The acceptable level of correlation coefficient would depend upon the degree of item homogeneity or item heterogeneity desired by the test constructor. In the case of an achievement test, the judgement of the subject matter expert must always play an important part in the selection and rejection of items.

Factor Analysis. It was shown that an observed score can be divided into two additive components, a true score and an error score. "In factor analysis it is assumed that the true score can be further subdivided into additive components due to various common factors and a factor specific to each test" (Baggaley, 1964, p. 98). Essentially,

> the principal concern of factor analysis is the resolution of a set of variables linearly in terms of (usually) a small number of categories of "factors". This resolution can be accomplished by the analysis of the correlations among factors which convey all the essential information of the original set of variables. Thus, the chief aim is to attain scientific parsimony or economy of description. (Harmon, 1960, p. 4)

Guilford (1965) has shown that many of the concepts of validity, e.g., predictive validity and multiple correlation principles, are explainable on the basis of factor theory.

The essential new step is to assume that the variance can be further broken down into independent additive components of common factor variance, specific variance and error variance. Communality is defined as the proportion of common factor variance in the test scores. The proportion of specific variance in a test is known as its specificity, which is symbolized by $\underline{S}^2$. Error variance is denoted by $\underline{e}^2$. In equation form, symbolizing total variance by 1.0,

$$1.0 = h^2 + s^2 + e^2$$

The specificity plus the error variance is called the uniqueness of a test, or

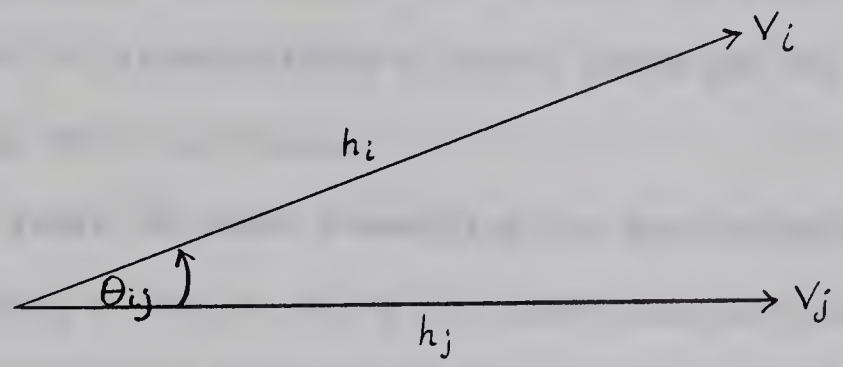
$$U^2 = s^2 + e^2$$

When factors are uncorrelated, factor loadings are always the coefficients of correlation between the respective factors and the variables that were factored. The correlation between two tests is the sum of the cross products of the common factor coefficients or factor loadings. In equation form, symbolizing the correlation coefficient by $\underline{r}_{ij}$ and the factor loadings by $\underline{a}_{ip}$ and $\underline{a}_{jp}$,

$$\hat{r}_{ij} = \sum_{p=1}^{n} a_{ip} a_{jp}$$

where $\underline{i}$ and $\underline{j}$ refer to tests and $\underline{p}$ denotes the $\underline{p}^{th}$ common factor.

A test score can be represented in a space as a point using the co-ordinates given by the factor loadings. The same geometry also holds for an item of a test. If an item was represented by the vector, $\underline{V}_i$ of length $\underline{h}_i$, and another item by vector, $\underline{V}_j$ of length $\underline{h}_j$, the relationship between the vectors can be shown as



where $\underline{\theta}_{ij}$ is the angular separation of the two vectors. In equation form,

$$\text{COS } \theta_{ij} = \frac{1}{h_i h_j} \quad r_{ij}$$

When several items have high loadings on a single factor, an indication is given of the internal consistency reliability index for these items. The items may be regarded as being comparable measures of the same hypothetical variable. If the correlations, in a correlation matrix that is to be factored are uniformly high, high factor loadings will result yielding a high internal consistency estimate of reliability. Also, when an item and a criterion both have high loadings on the same factor, it is an indication of validity for predicting that factor. "The correlation of a test with each common factor (a common factor loading) is its coefficient of validity for measuring that factor" (Guilford, 1954, p. 399).

Factor analysis can be used as an aid to solve the weighting problem for many different items forming a single test score. From an entire set of test items that have been factored, a common factor score can be calculated for each individual by the appropriate weighting of his scores on the original variables. The procedure may be used to calculate a single factor score derived from the first, and also the largest, factor or alternatively a factor score may be obtained for each individual on every factor.

Where there is doubt concerning the psychological homogeneity of the items forming a test or where the item-total correlation tends to be low, factor analysis may be used to divide the test into subtests, each of greater homogeneity. The assumption is that tests with high internal consistency are desired.

Relationship of Item to Test Score. Basically, in preparing a test one is concerned with the problem of selecting items so that the resulting measurement instrument will have certain specified characteristics. Flowers (1965) has suggested that given the item means, measures of item conformities and measures of item validities, it should be possible to assemble a test which could satisfy, within certain limits, a prescribed mean, standard deviation, reliability, validity, skewness and kurtosis. Although Gulliksen (1950 a) does not suggest that skewness and kurtosis should be completely ignored since they pose many as yet unsolved problems, he limits his suggestion to the possibility of selecting items to influence the test mean, variance, reliability and validity.

If transformations of test scores can be used to obtain scores with a specified mean, variance, and within certain limits the form of the score distribution, it is suggested that test construction procedures may profit when an emphasis is placed upon producing reliable and valid tests. "Tests composed of items answered correctly by about 50 percent of the group have a higher validity than tests composed of items that are easier or harder than 50 percent, but otherwise of the same type" (Gulliksen, 1950 a, p. 374). It has been shown by Gulliksen that the formula for calculating test validity does not show any direct relationship between test validity and item difficulty, but test validity however, does depend on the point-biserial item-criterion correlation.

Theoretically, the problem of selecting a subset of $\underline{k}$ items from a total group of $\underline{K}$ items as well as the problem of maximizing test validity for predicting any specified criterion has been solved. A completely accurate solution is obtained by using the interitem variance-covariance matrix to select the one subset of size $\underline{k}$ that has the highest validity. The procedure is very laborious. Gulliksen (1950 a) has reviewed several approximation procedures. If the complete interitem variance-covariance matrix and the item-criterion covariances are available, a maximum test validity may be obtained by solving for all multiple correlations or for all multiple correlations using a specified number of items.

The incompatabilities of attempting to construct a test with high validity and high reliability, where validity is the more important (Helmstadter, 1964; Ebel, 1965; Guilford, 1965), does not justify any

devaluation of high reliability as a goal in test construction. Reliability is essential to validity. However,

> validity is by far the most important criterion by which a test may be judged, for an objective, reliable, and well standardized instrument can still be completely useless unless the kinds of inferences which can legitimately be made from the test score are known (Helmstadter, 1964, p. 226)

Correction for Attenuation. When two variables are correlated, the errors of measurement if uncorrelated among themselves, lower the coefficient of correlation compared to that derived from perfectly reliable measures. It is possible but unlikely that a random change in score would make the correlation larger. McNemar (1962, pp. 153 - 154) has derived the correction for attenuation formula

$$r_{tt} = \frac{r_{xy}}{\sqrt{r_{xx}}\ \sqrt{r_{yy}}}$$

where r_{tt} is the correlation between perfectly reliable "true" scores on x and y.

r_{xy} is the correlation of actual scores on x and y.

r_{xx} is the reliability of the measure of variable x.

r_{yy} is the reliability of the measure of variable y.

A correlation coefficient corrected for attenuation may be regarded as

(a) the correlation between true scores in each of the two measures and

(b) the correlation between the two measures when each is increased to infinite length (and hence a reliability of 1.00). (Gulliksen, 1950 a, p. 101).

Gulliksen (1950 a) maintains that the "correction for attenuation" is not actually a "correction" but rather is an estimate of the correlation between a perfect test and a perfect criterion. Correction for attenuation is actually a special case of partial correlation with the errors $\underline{e}_x$ and $\underline{e}_y$ partialed out.

One practical application of the correction for attenuation is to determine what increase in reliability of test $\underline{x}$ or criterion $\underline{y}$, or both, would yield a more satisfactory value of the validity $\underline{r}_{xy}$. The equation is valuable in giving a quick indication of the utility of attempting to increase the test validity by increasing the test length. The correction for attenuation may thus be used to indicate the most profitable direction for further validation research. Another application suggested by Gulliksen (1950 a, p. 214) is in calculating a correction for the attenuation due to inaccuracy of reading essays. However, while correlation coefficients corrected for attenuation are of theoretical importance in the analysis of relationships in that consideration can be made for variable errors of measurement, they should not be reported with the implication that the higher coefficient has already been attained. Corrected $\underline{r}$'s cannot be used in prediction equations as prediction must necessarily be based on obtained, or fallible, rather than true scores.

A restriction on the size of the validity coefficient is imposed by the reliability of the criterion. "It is more important that the reliability of a criterion measure be known than that it be high" (Thorndike, 1949, p. 107) since the following formula

$$r_{tt} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

may be used to provide estimates of the validity coefficient of the fallible tests we are compelled to deal with. We have a more stable means of comparing test validities if something is known about the validity of the criterion.

If either r_{xx} or r_{yy} is underestimated, the corrected r_{xy} will be overestimated. If either reliability coefficient is overestimated, the corrected r_{xy} will be underestimated. A conservative approach would be to underestimate the corrected r_{xy}. Also, the method of estimating a reliability coefficient influences the value obtained. The question also arises as to which of the three main types of reliability coefficient is desirable in correcting for attenuation. "In general, the alternate forms approach is probably the best" (Guilford, 1965, p. 489).

It has previously been shown (Gulliksen, 1950 a, p. 382) that

$$r_{xy} = \frac{\Sigma s_g r_{yg}}{\Sigma s_g r_{xg}} ,$$

which is the ratio of the average item validity indices to the average reliability indices. Here we have what Loevinger (1954) has called the attenuation paradox. The empirical validity of a test is decreased as the internal consistency of the test, measured by the item-total test score correlation, is increased. An increased internal consistency may also increase the external criterion correlation but beyond a certain

$$\frac{\partial^2}{\partial x^2}$$

point increase in internal consistency begins to eliminate relevant variance, thereby reducing the test-criterion correlation.

Each variable in a factor analysis is commonly treated as though it contains three components: common, specific and error variance. The variance components are illustrated by

$$1.0 = h^2 + s^2 + e^2$$

where $\underline{h}^2$ is the common variance (communality), $\underline{s}^2$ is the specific variance and $\underline{e}^2$ is the error variance. Essentially, the unique variance which is not common to the other variables is removed through estimation of communalities before the analysis is begun, or by selecting a small number of common factors after the analysis has been completed. The reproduced correlations are than attributable to only common factor variance.

CHAPTER V

REVIEW OF ITEM SELECTION PROCEDURES

Exact procedures have been developed for selecting items to form a test by using complete regression systems. Some procedures allow positive and negative weights for each item whereas other methods designate selection or rejection of an item by a weight of one or zero. Because such methods were regarded as too laborious computationally for practical purposes, several approximation techniques have been devised.

Weighting

When a single score is to be derived from a weighted sum of items, one is faced with the problem of determining the appropriate method of combining these scores. One solution is to select the items for a test and then use multiple correlation procedures to determine the optimal weighting system for each item in the test to predict a selected criterion. An alternative approach is to select items by using a step-wise regression solution. Theoretically the method of using weights is the most suitable for accurate prediction but it tends to increase to a considerable extent the time and effort involved in determining an individual's score. Gulliksen suggests that some approximation to multiple correlation is to be preferred to the exact method, when selection is to be made from many variables, since "for practical purposes, simple integral approximations to the exact multiple weights will usually give a satisfactory composite score" (Gulliksen, 1950 a, p. 356). Douglas and Spencer (1923) concluded that it made very little difference in the

ultimate outcome as to what weights are assigned to each measure. They found, for a number of tests, that scores obtained with unit weights correlated .98 to .99 with the same scores obtained through use of optimal item weights. It may therefore be concluded that using fractional weights rather than integral weights for different items in a typical test will not prove significantly more valuable in arriving at a total test score. "The gain in predictive efficiency achieved by the use of ultra-refined techniques of item analysis in preference to relatively crude methods would appear to be nominal at best" (Rozeboom, 1966, p. 519).

Regression Procedures

Various procedures have been reported which enable a test constructor to maximize test validity by selecting individual items from a pool of items. If a criterion is available, and we desire to weight the items in such a manner that the composite score will have the highest possible correlation with the criterion, the method of multiple correlation is the one to use (Gulliksen, 1950 a). The above procedure has been extended by Horst (1961) to include a set of $\underline{n}_1$ predictor variables and a set of $\underline{n}_2$ criterion variables. If the $\underline{n}_1 + \underline{n}_2$ variables for the same individuals are available, a linear combination of the predictor variables and a linear combination of the criterion variables can be calculated which will yield the highest possible correlation between the composites. Gulliksen maintains that multiple correlation methods give the best weights for predicting the criterion but "simple integral approximations to these weights will usually give a composite score

that correlates almost as well with the criterion" (1950 a, p. 330). The above procedures involve finding the "best" weighting system for a given set of items whereas a step-wise regression procedure allows one to select an item at a time which will result in an ordered selection of the subset of k items that best predicts the criterion. However, "the precise method of weighting is not important unless we are dealing with relatively few tests that are not highly correlated" (Gulliksen, 1950 a, p. 327).

The procedure for predicting an external criterion by multiple correlation is outlined by Gulliksen (1950 a) in Theory of Mental Tests. Several methods of selecting items for a test by approximations to multiple correlation have been published.

Approximation methods to multiple correlation have been developed which are used to assemble a collection of items whose composite score would have maximum validity. Horst (1936) proposed a method which takes into account the intercorrelations of the items as well as their correlations with the criterion. Other closely similar procedures have been described by Richardson and Adkins (1938), Toops (1941), Wherry and Gaylord (1946), Gleser and DuBois (1951), Horst (1956), and Horst and MacEwan (1956, 1957). Lubin and Osburn (1957) reported a technique of pattern scoring of test items for the prediction of a quantitative criterion. Osburn and Lubin (1957) have worked with a method whereby test scoring techniques can be evaluated to see if they have maximum validity. A less laborious, though analogous, procedure than that of Gleser and DuBois (1951) has been developed by Webster (1956). Webster's non-parametric method will yield dependable results for dichotimized

items when N (observations) is large.

Canonical Correlation

A statistical procedure, seldom mentioned in references dealing with test theory and item selection, known as canonical analysis may be an appropriate technique that should be applied to the general area of item selection. Canonical analysis is another approach to multivariate analysis. In canonical analysis the linear combination of the dependent variables which are the most predictable from the best linear combination of the independent variables is found (Cooley and Lohnes, 1962).

Before the advent of large and fast computers, canonical analysis was far too time consuming and involved to be practical. However, present facilities are available to handle the tremendous number of calculations involved. Canonical correlation techniques could be used to find regression weights for the items and criterion variables. Although items with low regression weights can, in subsequent analyses, be omitted from a test, which is in a sense a means of selecting items, the procedure is not in fact well suited for the problem of selecting items. The merit lies in weighting the variables after the selection of items has been completed. However, when many items are used to construct a test, the particular weighting system is not too important. An important aspect of using canonical correlation techniques is that a multi-dimensional criteria space is considered in selecting weights for the predictors. However, while the linear weighting system applied to the items and criteria may tell us something about the items, in many situations the user would like to specify his own linear combination

of criteria.

Factor Analysis

Factors can be conceived of as the principles of classification or dimensions that allow the test constructor to reconstruct the properties of the material being considered rather than relying on subjective preference, intuition, or common sense (Eysenck, 1966). Factor analysis "enjoys its greatest justification as an exploratory technique, by which the variables under consideration all enjoy a reasonably well-rationalized, but not necessarily certain, probability of belonging to the scientific domain of interest, and the structure of which is essentially unknown" (Kaiser, 1966, p. 361).

Items may be sampled after a simple structure factor loading matrix has been calculated. The items having the highest loadings in each factor are defined as the best measures of that factor. Several tests can be formed. Each test or subtest is constructed by including those items with the highest loadings in each factor. Since the constructed tests will be mutually orthogonal, the common procedure is to form a battery of tests (Horst, 1965, 1966).

Scale Analysis

In references relating measurement to scale analysis (Lingoes, 1963; Horst, 1965), Guttman's name is frequently mentioned. Guttman's (1955) concept of a perfectly scalable set of items was based upon the notion that all persons marking an item with a given preference value would also mark all items of greater preference value. Thus, for any particular set of item difficulties, a person getting a more difficult

item correct would also get all the easier items correct. The resultant matrix has been called a perfect simplex. If a perfectly homogeneous set of items with resulting perfect retest reliability is available, one item is of as much value for measurement purposes as the entire set of items. However, as a result of varying item difficulties and measurement errors, the ideal test item is not available.

The concept of a "universe of content" in relation to item construction and selection in Guttman's scalogram method has recently been extended by Lingoes (1963) who presented a completely objective and empirical procedure for selecting dichotomous items which meet the Guttman scaling criteria in multiple dimension situations. Lingoes' method involves

> selecting an item from the set to be analysed, finding that item among the remaining items which is most like it and having the fewest errors, determining the number of errors between the candidate item and all of its predecessors, and finally, applying a statistical test of significance to adjacent item pairs. ... All items are forced into a positive manifold and monotonicity of item marginals is insisted upon (1963, p. 502).

No references or examples concerned with the above procedure have been found by the writer in a review of test construction procedures. The technique appears to offer a new apporach to selecting items but it will have to be tested empirically prior to any conclusive decision regarding usefulness.

CHAPTER VI

THEORETICAL DEVELOPMENT OF THE SELECTION TECHNIQUE

For selection purposes the pool of items may be considered to be responses made by examinees to the items. With knowledge of the response to an item and the associated correct response, it is possible to calculate representative summary data for the items. A common analytic representation of the relationships between items is given by a correlation matrix and the corresponding array of item means.

Although information about the relationships among the variables in a multiple set is summarized by means of a correlation matrix calculated from multiple measures, the problem of interpretation and summarization is encountered. It would be difficult for a person to relate fully all variables and subsequently provide an interpretation of all the relationships. Part of the difficulty arises because of the over-lapping nature of the variables.

Factor analysis can be used to approximate the original relationships among variables in terms of a smaller number of basic constructs called factors. The dimensionality of the factor matrix will be, in most cases, much smaller than the rank of the correlation matrix. Concern here is with the minimum number of factors and not with their interpretation.

The problem of interpretation can best be handled by the use of Thurstone's (1947) principle of simple structure. If the simple structure criterion is used for finding the factor loading matrix corresponding to a correlation matrix, it should be relatively easy to

identify meaningfully the factors. Since interpretation rests heavily upon a pure definition of items and criteria, the problem of meaningfulness is the responsibility of the user. However, the problem of defining an analytic criterion for selecting items is independent of the factor interpretation problem. A solution, based upon psychometric principles, is outlined below.

Factor Analysis and Prediction

"Many methods are available for predictor selection, but in general, these have not used the factor analytic techniques and are deficient in that they capitalize on chance error" (Horst, 1965, p. 22). Only recently, with the advent of electronic computers, have factor analytic techniques been applied to the problem of predictor selection. Although it has been claimed that multiple regression provides the best theoretical answer to the problem of developing a test to predict a single criterion (Gulliksen, 1950 a), Horst maintains that

> it is of some interest, however, to see not only how the classical methods of least square multiple recti-linear prediction can be brought into the general framework of factor analysis, but also how these classical methods can be modified, and perhaps even improved by formulating the problems in terms of the models and objectives of factor analysis (Horst, 1965, p. 540).

A distinct advantage of factor analytic techniques is that estimates of error variance may be introduced into the solution.

Proposed Selection Technique

The item-selection method proposed here begins with a principal axis factor analysis of the matrix of intercorrelations calculated from predictor and criterion raw score data. One solution to the problem

of deciding how many orthogonal factors to extract has been proposed by Kaiser (1960) who recommends using only those factors with corresponding eigenvalues greater than one to define the common factor space. The rule applies only if unities are used in the diagonal of the correlation matrix. If $\underline{h}_j^2$ represents the communality of variable $\underline{j}$, the remaining 1.00 - $\underline{h}_j^2$ variance of item $\underline{j}$ is the unique component or residual variance. The resulting factor solution can be used to calculate $\hat{\underline{R}}$, a reproduced correlation matrix, that is, an approximation of the original correlation matrix $\underline{R}$. A measure of the lack of fit between the obtained factor model for the domain and the observed relations among the variables is given by the difference between $\hat{\underline{R}}$ and $\underline{R}$. It is assumed that the true variance is free from error or random variation. A simple structure transformation of the principal axis factor loading matrix is often required for purposes of interpretation. A major objective in using factor analysis is to be able to identify and eliminate as much unsystematic variance as possible. Secondly, it is highly desirable to have projections of item variables on orthogonal factors to represent the idealized dimensions. A convenient rotational procedure to achieve a simple structure approximation is recommended by the writer prior to using the proposed selection method. A previous and still frequently used application of factor analysis is to select subsets of variables such that each of the simple structure factors will be adequately represented by the subset. This is essentially the common procedure in selecting tests to form a test battery.

Consideration must also be made of criterion measures required to

establish psychological meaningfulness for the test under construction. It is argued that even though ultimate criteria may not be readily available, the test constructor must have some notion of what relatively unitary skills, aptitudes or abilities are required. An investigator must locate measures which, through the use of a linear weighting system can be made to approximate the ultimate criterion. When dealing with several criteria, attempts are generally made to combine them into a single criterion measure or to use each as a single criterion score. The use of several independent criteria provides a simplified solution but yields less information than a composite criterion. As in the case of predictors, it is possible to identify the common factor variance of the separate criteria by means of factor analysis. Those criteria with high factor loadings could be selected to represent the best measures of the factors in the criterion space. The use of basic statistics to determine which criterion to use does not, however, deal directly with the fundamental problem of relevance. Some decision must be made by the investigator, or group of "experts", as to the relevance of each criterion to the ultimate criterion.

When questions of concurrent and predictive validity arise, the first concern is with finding a suitable criterion. The establishment of content validity requires no measurable criterion since to assign content validity to any test, is, in essence, to compare idealized course content with examination content. A general impression formed by the writer after reading references concerned with various aspects of validity, e.g., Technical Recommendations (1954), Cronbach (1960) and Anastasi (1961), is that little, if any, concern is given to the notion of predictive test

validity until after a test has been constructed. This is not to say that there is no thought given to defining the end product in terms of some set of criteria. However, the implicit under-lying set of rules used to develop a test should first be made explicit and formally stated. This is done at times in constructing items when "objectives" and "content" areas are used to specify types of required items. Important preliminaries to the construction of any test should be a specification of ultimate criteria and some attempt at criterion development.

The proposed item selection technique is applicable to the problem of selecting items where the item variables can be described in terms of a known factor structure. The classical situation where the test score is a linear function of the item responses will be considered. An item response is to be used as a predictor of a criterion variable. Sampling of questions is not considered. Selection of items is restricted to a design problem. The proposed item selection approach is based upon the assumption that the items and criteria have a known factor structure with a comparatively small number of common factors. This is a departure from other selection models in that a reduced matrix of factor coefficients is used as a starting point. The rank of the reproduced correlation matrix will be considerably smaller than the order of the original correlation matrix.

General Description of the Selection Procedure

The factor analytic procedure used by the writer for obtaining a factor matrix from the correlation matrix, composed of items and criteria, was a principal axis factor analysis using the Householder method (1938) with unities in the diagonal of the correlation matrix. A primary concern

at this stage was to establish the number of significant orthogonal factors, according to Kaiser's criterion (1960) of retaining those factors with eigenvalues greater than one. Kaiser's (1958) varimax criterion was applied to rotate the principal axis factor matrix to simple structure. The procedure outlined above has been used commonly in the application of factor analytic techniques to the isolation and identification of a limited number of hypothetical variables underlying a group of observed variables.

A transformation matrix was used to rotate the factor matrix such that the hypothetical goal test vector, specified by the test constructor, and the first factor were collinear. A geometric representation of a three-dimensional orthogonal factor space is presented in Figure 1. Factors represented by axes I, II and III give the geometric basis for determining the location of the goal vector (GV). By assigning relative weights, x_i, to each factor, the location of GV in the item and criterion space is specified. The axes are each rotated $\emptyset^{o}$, as illustrated, to position factor I collinear with GV. Axes I', II' and III' are now the frame of reference for the orthogonal factor space. The loadings on factor one then represent the correlation of each item, as well as the criteria, with the goal test. With the direct relationship of an item to the goal test specificied, selection of items to construct a test is initiated.

In Figure 2 the relative positions of five items in the factor space, the loadings on GV and the three axes are illustrated. Each item is numbered according to the order of selection in constructing a test. Items 4 and 5 would not be selected because they do not meet certain

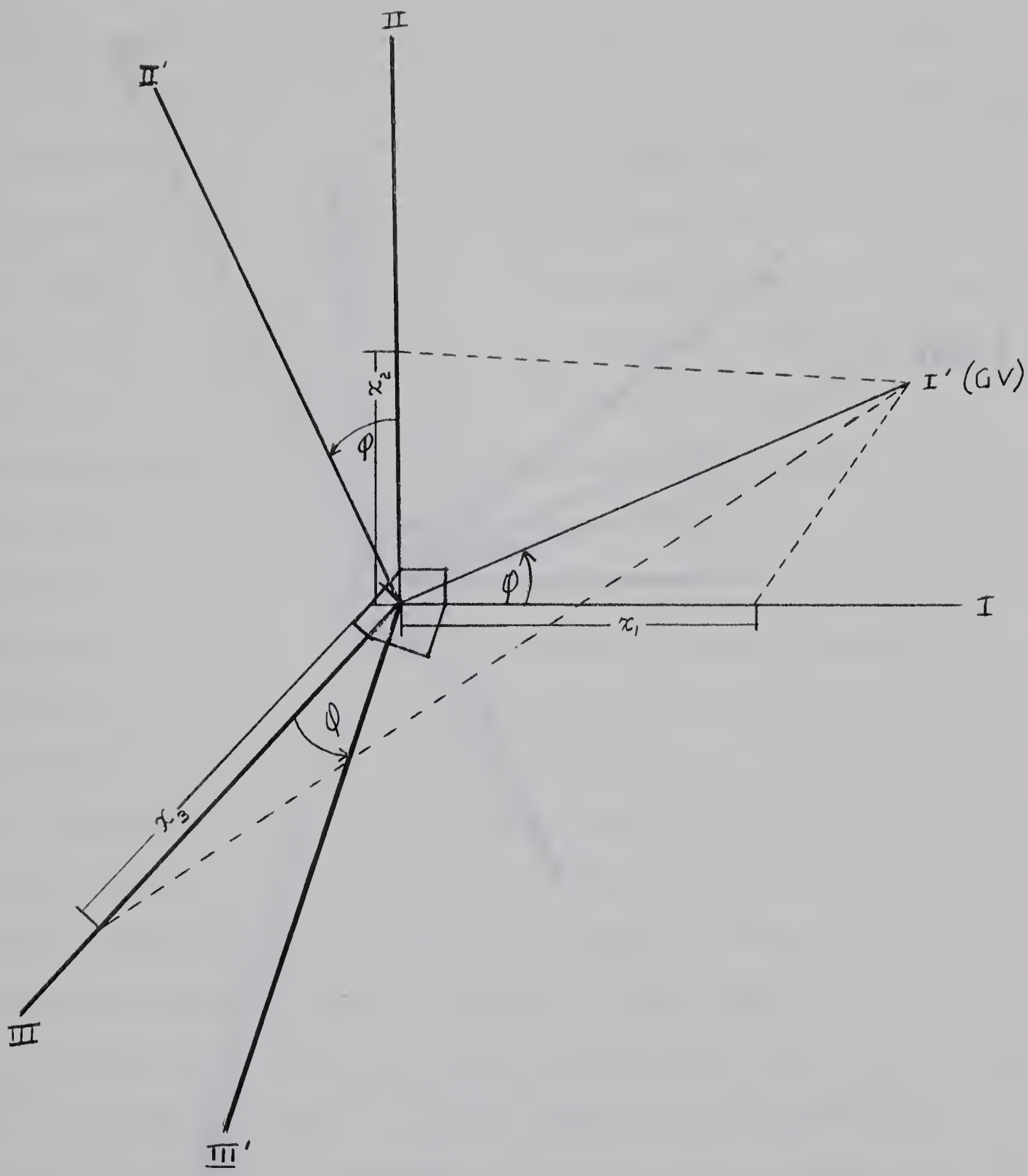


Figure 1. Position of the Goal Test Vector in the Common Factor Space

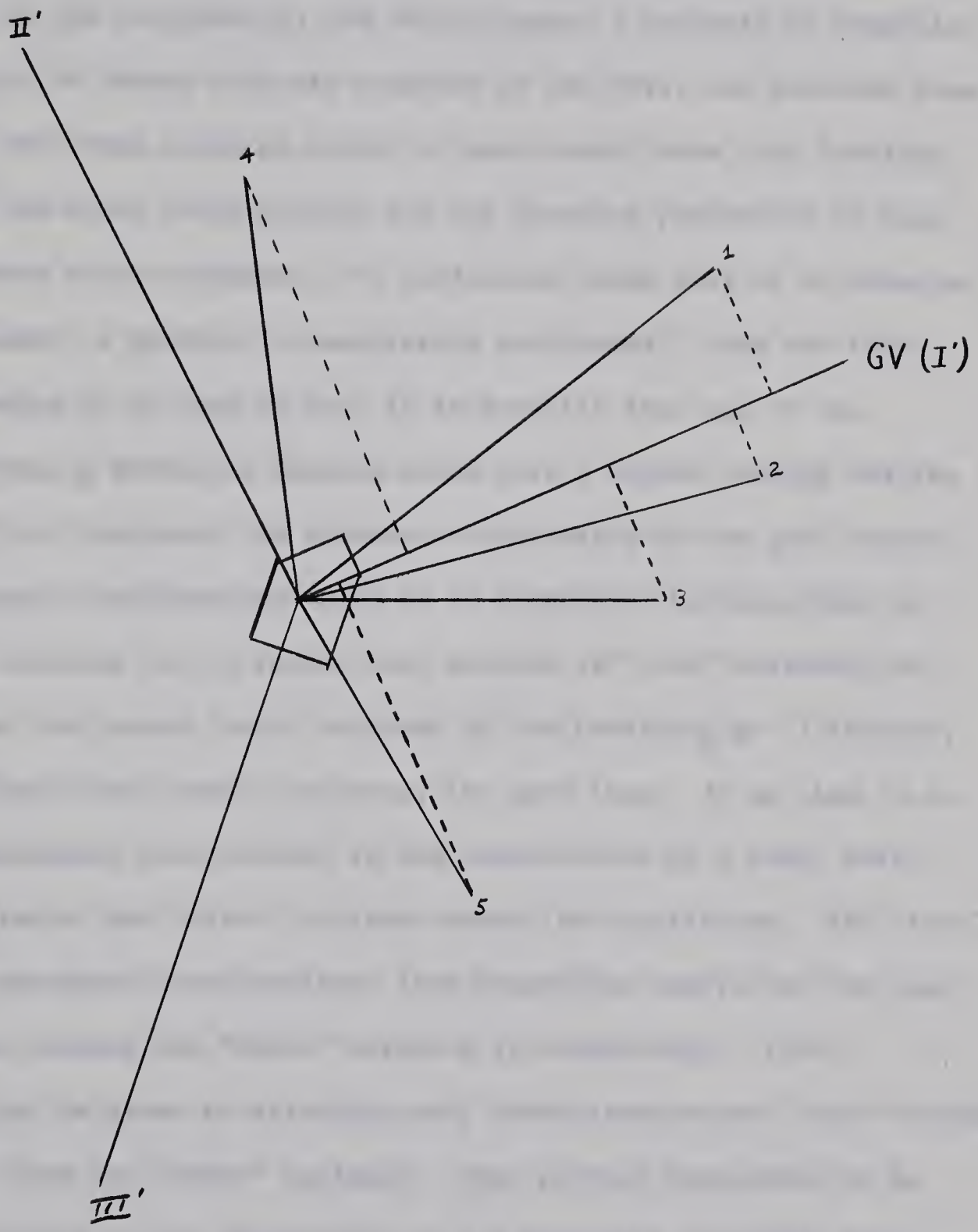


Figure 2. Relative Positions of Item Vectors in the Common Factor Space

specifications that will be explained later. The first item selected would have the highest loading on factor one. In sequence, the second item selected would load second highest on the first factor. From knowledge of the locations in the factor space, a centroid or composite vector would be formed that was composed of the first two selected items. The first two items selected would, in most cases, have high loadings on factor one which would account for the greatest proportion of each item's common factor variance. If additional items were to be selected in this manner, a problem is immediately manifested. When the first factor loading of an item is low, it is possible that one of the second through m extracted factors would have a higher loading and the item would not represent the intended relationship to the goal vector. A more general consideration would be to determine the proportion of variance accounted for by factor one, denoted as "true" variance, as compared to the common factor variance in the remaining m - 1 factors, here referred to as "error" variance, for each item. If an item is to make a significant contribution in the construction of a test, more "true" variance than "error" variance should be contributed. The "true" variance represents characteristic item properties desired by the test constructor whereas the "error" variance is undesirable. Thus, consideration is given to selecting only those items whose "true" variance is greater than the "error" variance. That portion considered to be "error" variance in one application of the selection procedure may represent variance components of other tests orthogonal to the goal test.

The angular departure and the correlation relationship of an item to the goal test vector should be considered. When an item vector

deviates more than 45° from the goal vector, the variance contributed to the goal vector will be less than that to tests orthogonal to the goal vector. Thus, the item would not be considered for selection. An item correlation of less than .300 with the goal vector would not be considered significant because the additional variance contributed by this item would not appreciably add to the specification of the goal test. A final item characteristic should be that the loading of an item on factor one be greater than or equal to .300 in order to be considered worthy of selection. For the reasons given above, items 4 and 5 in Figure 2 would not be selected.

After consideration has been given to the various restrictions to be imposed prior to selecting an item the selection of items continues until the desired number of items have been selected, until there are no items remaining in the pool of items that meet the imposed conditions, or until the test being constructed deviates in composition from the prescribed tolerance limits.

The procedures suggested by the writer in the above presentation are not intended to provide restrictions upon the use of the factor analytic item-selection algorithm. Many variations would be possible if various types of factor analysis such as the square root procedure, the maximum likelihood solution or the alpha factor analytic method replaced the principal axis solution. In addition, the equimax or quartimax criterion could be used in preference to the varimax criterion example. Several parameters were presented above to suggest a maximum acceptable angular displacement and a minimum significant correlation coefficient. These values are, in the writer's opinion, merely plausible suggestions

and may thus be varied to meet the individual test constructors' specifications. It is intended that each test constructor will, with minimal effort, be able to modify the test parameters to best represent the desired characteristics in the constructed test.

Mathematical Description of the Selection Procedure

Mathematically the procedure for selecting items can be stated as follows.

Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \\ \hline c_{11} & c_{12} & \cdots & c_{1m} \\ \vdots & \vdots & & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{km} \end{bmatrix}$$

where the elements a_{ij} are the factor coefficients of n predictor variables (items in our case) on m orthogonal factors and c_{rj} are the factor loadings for the k criteria on the m orthogonal factors.

Since the m factors of the space A_m span the space and are, from the users' point of view psychologically meaningful, the object test

taken as a linear combination of the m factors will also be contained within the space $\underline{A}_m$. That is, the object test will be contained within the same space as the common parts of both item and criterion factors. The exact linear combination of the m factors required to define the object test vector may be specified as

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{bmatrix}$$

which has as elements the relative weighting system to be applied to the m orthogonal factors.

In order to determine the perpendicular projections of each item vector upon the object test vector, it is desirable to place any one of the m orthogonal factors collinear with X. Arbitrarily, the first vector may be selected and positioned by defining an appropriate transformation matrix T to be applied to the matrix A.

Specifying the normalized vector X as t and appending it to the matrix A such that

$$A = \begin{bmatrix} a \\ --- \\ c \\ --- \\ t \end{bmatrix},$$

it is required that the transformation matrix $\underline{T}$ be such that

(a) $\underline{AT} = \underline{S}$ where $\underline{r}_{s_1 t} = 1$ and $\underline{r}_{s_r t} = 0$ for $\underline{r} = 2, 3, \ldots \underline{m}$ $(\underline{r} \neq \underline{t})$

(b) and that $\underline{T}\,\underline{T}' = \underline{I}$, that is $\underline{T}$ is an orthonormal transformation matrix performing an orthogonal transformation on $\underline{A}$.

Such a transformation matrix may be generated in a number of ways, perhaps the simplest being as follows.

Let

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_1 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{m-1} & x_m & \cdots & x_{m-2} \\ x_m & x_1 & \cdots & x_{m-1} \end{bmatrix}$$

and apply the Gram-Schmidt orthonormal process (Hohn, 1964, pp. 264 - 267) to $\underline{X}$, starting with the column vector $\underline{X}_1$ in forming

$$T_1 = \frac{X_1}{\left| X_1 \right|}$$

and for the $\underline{r}^{th}$ column vector of $\underline{T}$, $\underline{T}_r$ is given by

$$T_r = \frac{X_r - \sum_{k=1}^{r-1} (T_k' X_r)\, T_k}{\left| X_r - \sum_{k=1}^{r-1} (T_k' X_r)\, T_k \right|} \quad \text{where } \underline{r} = 2, 3, \ldots \underline{m}$$

yields the remaining column vectors of $\underline{T}$.

The $\underline{n} + \underline{k} + 1$ vector of $\underline{A}$ is defined as the vector $\underline{T}_1'$. The matrix $\underline{T}$ may now be used to rotate the matrix $\underline{A}$ such that the vector with projections $\underline{S}_1$ of $\underline{S}$ is collinear with the column vector $\underline{T}_1$.

Let

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \\ \hline c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{km} \\ \hline t_{11} & t_{21} & \cdots & t_{m1} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mm} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nm} \\ \\ \vdots & \vdots & & \vdots \\ \\ s_{g1} & s_{g2} & \cdots & s_{gm} \\ \\ s_{h1} & s_{h2} & \cdots & s_{hm} \end{bmatrix}$$

$\underline{A}$ $\underline{T}$ = $\underline{S}$

where $\underline{g} = \underline{n} + \underline{k}$ and $\underline{h} = g + 1$. Since $\underline{T}$ is orthonormal $s_{h1} = 1.000$ and s_{h2} through s_{hm} are equal to 0.000.

It would be desirable that the sums of squares, $\underline{SS}$, be as follows

$$SS_r = \sum_{j=1}^{n} S_{jr} \qquad (r \neq 1) \text{ be such that}$$

$$SS_r \geq SS_{j\ r+1} \qquad (r \neq 1).$$

If this condition were attained the $\underline{m}$ - 1 remaining vectors of $\underline{S}$ would account for decreasing importance in redefining the original space $\underline{A}_m$. Since the first vector $\underline{S}_1$ of $\underline{S}$ represents the object test, in the same sense the remaining $\underline{m}$ - 1 vectors of $\underline{S}$ represent 'other tests' orthogonal to the object test. Since an item does not contribute solely to a single object test but also to all other tests orthogonal to it in the space $\underline{A}_m$ (except in the case of the item being perfectly orthogonal to one or more $\underline{S}_r$, $r \neq 1$), the contributors of the item to the remaining orthogonal tests should also be assessed.

Therefore it is desirable to arrange our transformation such that the orthogonal tests account for decreasing amounts of the remaining common item-criterion variance. Thus, we now adjust the $\underline{S}_2$ through $\underline{S}_m$ column vectors of $\underline{S}$ to have decreasing amounts of variance accounted for by each factor. Let

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1p} \\ f_{21} & f_{22} & \cdots & f_{2p} \\ \vdots & \vdots & & \vdots \\ f_{h1} & f_{h2} & \cdots & f_{hp} \end{bmatrix} \equiv \begin{bmatrix} s_{12} & s_{13} & \cdots & s_{1m} \\ s_{22} & s_{23} & \cdots & s_{2m} \\ \vdots & \vdots & & \vdots \\ s_{h2} & s_{h3} & \cdots & s_{hm} \end{bmatrix}$$

where $\underline{p} = \underline{m} - 1$. Considering again an orthogonal transformation matrix

$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

E such that

$$(F\ E)'\ F\ E = \lambda,$$

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \lambda_{m-1},$$

it can be seen that E are the eigenvectors of F F' and λ their associated eigenvalues, i.e.

$$E'\ F'\ F\ E = \lambda$$

$$F'\ F = E\ \lambda\ E'$$

The matrix E will provide a transformation for which the m - 1 remaining vectors of F will be of decreasing importance in terms of accounted variance. The m - 1 remaining vectors may be considered 'concomitant tests' orthogonal to the goal test each of decreasing importance. Multiplying so that

$$F\ E = D$$

will yield sums of squares of decreasing order for the D matrix. The elements of D are appended to the first column vector S_1 of S to produce

$$\left[\begin{array}{c:ccc} S_1 & D \end{array}\right] = \left[\begin{array}{c:ccc} s_{11} & d_{11} & d_{12} & \ldots & d_{1p} \\ s_{21} & d_{21} & d_{22} & \ldots & d_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ s_{h1} & d_{h1} & d_{h2} & \ldots & d_{hp} \end{array}\right] \equiv B$$

The object test vector is defined as the row vector

$$\left[\begin{array}{cccc} 1.000 & 0.000 & 0.000 \ldots & 0.000 \end{array}\right]$$

Our task now is to select predictor variables from the n row vectors of

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \\ \vdots & \vdots & & \vdots \\ b_{h1} & b_{h2} & \cdots & b_{hm} \end{bmatrix}$$

such that the centroid of the selected vectors will be nearly collinear with the object vector.

Appendix A contains a flow chart of the item-selection algorithm.

Criteria for Item Selection

To this point the theory indicates that a precise formulation of a goal test can be specified within the space $\underline{A}_m$ and in such a manner has to have psychometric meaning. The selection of items to approximate the object test is now required.

From the geometry of the space $\underline{A}_m$ containing $\underline{n} + \underline{k} + 1$ vectors, the following selection procedures appear reasonable:

(a) At any stage of selection the correlation, $\underline{r}_{gc}$, between the goal test and the composite vector should be a maximum where $\underline{g}$ represents the goal test and $\underline{c}$ the item composite approximation to $\underline{g}$. Since $\underline{g}$ is of unit length ($h_g^2 = 1.000$) while $\underline{h}_c^2 \leq 1.000$ and in the

empirical case $\underline{h}_c^2 < 1.000$, and since

$$r_{gc} = h_g\ h_c \cos \theta$$

$$r_{gc} = h_c \cos \theta$$

for a given $\underline{\theta}$ selecting items in terms of decreasing communality will produce a decreasing $\underline{r}_{gc}$. Similarly, for a fixed $\underline{h}_c$ selecting on the basis of largest to smallest $\underline{\theta}$ will have the same effect. Selecting on both $\underline{h}_c$ and $\underline{\theta}$ results in the same trend. Thus, by this procedure a negatively sloped function for $\underline{r}_{gc}$ is to be expected.

(b) Select all items that have

$$b_{il}^2 > \sum_{r=2}^{m} b_{ir}^2$$

which is equivalent to

$$(h_i^2 - b_{il}^2) < b_{il}^2 \qquad \text{(SC2)}$$

Since

$$\cos \theta_{ij} = \frac{1}{h_i\ h_j}\ r_{ij}$$

can be reduced to

$$h_i \cos \theta_{ij} = b_{il}$$

because of the unique properties of the object test vector, the above equation relating communality to the first factor variance can be written as

$$(1 - \cos^2\theta_{ij}) < \cos^2\theta_{ij}$$

which can be transformed to

$$0.5 < \cos^2\theta_{ij}.$$

Thus, given the condition

$$b_{i1}^2 > \sum_{r=1}^{m} b_{ir}^2,$$

the restriction

$$0^{\circ} \leq \theta_{ij} < 45^{\circ}$$

is imposed.

If an item $\underline{h}_i^2$ was very low, even though the above condition was met, the item would be included. This would not be a fully adequate criterion since if $\underline{h}_i^2$ was low the associated item should not remain in the pool or space $\underline{A}_m$. The specification of a minimum acceptable value SC1 of $\underline{b}_{i1}$ would solve the communality problem presented above.

(c) Termination of the selection method is dependent upon additional stop criteria specified by the user. Selection of items will be discontinued when

1. the correlation between the composite vector and the object vector is less than SC3 or
2. a maximum angular departure of the composite vector from the object vector is greater than SC4, or
3. there are no further items remaining after meeting the conditions imposed by SC1 and SC2, or
4. the number of items desired by the test constructor have been selected.

Each test constructor must provide values for SC3 and SC4 which act as a "stop criterion" for the item-selection process.

Validity and Reliability Estimates

The test validity for n selected items can be estimated by calculating a correlation coefficient to determine the relationship between the composite test vector and the goal test vector. It is assumed that the goal test vector represents the composite criterion. Classical test validity methods presented in the review of the literature on measurement theory are not fully appropriate in the present situation although the notion of correlating a series of predictors with a criterion score is retained. The defined test validity coefficient most appropriate in relation to the proposed method is the correlation expressing the relationship between the composite test vector and the goal test vector.

A centroid of any n selected items locates the composite test vector in the item and criterion space. The centroid would have m co-ordinates

$$\left[\frac{1}{n} \sum_{k=1}^{n} a_{k1}, \frac{1}{n} \sum_{k=1}^{n} a_{k2}, \ldots, \frac{1}{n} \sum_{k=1}^{n} a_{km} \right]$$

By restricting the number of factors to m, a limitation is imposed upon reproducing the original correlation coefficient between two vectors. Since the validity coefficient defined above is based upon the "goodness of fit" of one vector to the location of another vector, the validity coefficient may be thought of as a coefficient of reproducability. The validity coefficient is

$$r_{val} = h_c \cos \theta_{cg}$$

where θ_{cg} is the angle between the composite test vector and the goal test vector and h_c is the length of the composite test vector.

A relatively simple procedure is available for calculating a test reliability coefficient. Cronbach (1951) has shown that one of the Kuder and Richardson (1937) formulas gives the mean of the correlations resulting from all possible ways of splitting a given test into two halves and that it gives the proportion of first-factor variance extracted from the intercorrelations of the test items. Thus,

$$\frac{1}{n} \sum_{j=1}^{n} a_{j1}^2$$

yields an estimate of test reliability.

The internal consistency reliability coefficient can be considered from two points of view. If the projections of the item vectors are on the goal test vector, one estimate is available regarding proportion of variance associated with a given criterion. However, if projections of item vectors are onto the composite test vector, the internal consistency coefficient is then truly an estimate of the constructed test's internal consistency variance. Depending upon the interpretation desired, either coefficient would be suitable. Thus, it may be advantageous to calculate both internal consistency reliability coefficients and subsequently label each according to the vector representation. Projections upon the composite test vector are easily found by using the normalized centroid locations of the composite test vector as a transformation matrix to rotate the item vectors such that the composite test vector and factor one are collinear.

A matrix A of the factor loadings for all n selected items is rotated by V, a vector composed of the normalized centroid locations of the composite test vector to form C which is a column vector containing the projections of the item vectors onto the composite test vector.

If the composite test vector was normalized, a procedure analogous to the correction for attenuation of a correlation coefficient would result. The validity coefficient would then be

$$r_{val} = \cos \theta_{cg}$$

Worked Example of the Selection Technique

The following is a sequential step by step worked example of the proposed analytical method for selecting items. It is assumed that the items have been previously written and administered to a large sample of subjects. We now start with the factor pattern of the items and criteria where

$$A = \begin{bmatrix} 0.440 & 0.320 & -0.080 \\ 0.460 & 0.190 & -0.610 \\ 0.840 & 0.030 & 0.020 \\ 0.640 & 0.460 & 0.190 \\ 0.560 & -0.600 & 0.340 \\ 0.550 & -0.390 & 0.220 \\ 0.590 & -0.100 & 0.120 \\ \hline 0.450 & 0.480 & -0.700 \\ 0.400 & -0.500 & 0.100 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \text{Items} \\ \\ \\ \\ \text{Criteria} \end{matrix}$$

The assigned weights are, respectively,

$$\begin{bmatrix} 5.000 & 2.000 & 1.000 \end{bmatrix}$$

which is the column vector $\underline{X}_1$. We now form $\underline{X}$

$$\begin{bmatrix} 5.000 & 2.000 & 1.000 \\ 2.000 & 1.000 & 5.000 \\ 1.000 & 5.000 & 2.000 \end{bmatrix}$$

and apply the Gram-Schmidt orthonormal process to X, starting with the column vector $\underline{X}_1$

$$T_1 = \begin{bmatrix} 0.913 \\ 0.365 \\ 0.183 \end{bmatrix}$$

which is used as a basic reference point to calculate

$$T = \begin{bmatrix} 0.913 & -0.185 & -0.364 \\ 0.365 & -0.030 & 0.930 \\ 0.183 & 0.982 & -0.040 \end{bmatrix}$$

The matrix $\underline{T}$ is now used to rotate the matrix $\underline{A}$ ($\underline{AT} = \underline{S}$) such that the column vector represented by $\underline{S}_1$ of $\underline{S}$ is collinear with the column vector $\underline{T}_1$.

$$\begin{bmatrix} 0.440 & 0.320 & -0.000 \\ 0.460 & 0.190 & -0.610 \\ 0.840 & 0.030 & 0.020 \\ 0.640 & 0.460 & 0.190 \\ 0.560 & -0.600 & 0.340 \\ 0.550 & -0.390 & 0.220 \\ 0.590 & -0.100 & 0.120 \\ \hline 0.450 & 0.480 & -0.700 \\ 0.400 & -0.500 & 0.100 \\ \hline 0.913 & 0.365 & 0.183 \end{bmatrix} \begin{bmatrix} 0.913 & -0.185 & -0.364 \\ 0.365 & -0.030 & 0.930 \\ 0.183 & 0.982 & -0.040 \end{bmatrix} = \begin{bmatrix} 0.504 & -0.169 & 0.141 \\ 0.378 & -0.690 & 0.034 \\ 0.781 & -0.136 & -0.279 \\ 0.787 & 0.055 & 0.187 \\ 0.354 & 0.248 & -0.776 \\ 0.400 & 0.126 & -0.572 \\ 0.524 & 0.012 & -0.313 \\ \hline 0.458 & -0.785 & 0.311 \\ 0.201 & 0.039 & -0.615 \\ \hline 1.000 & 0.000 & 0.000 \end{bmatrix}$$

Items

Criteria

Object Test

A T = S

The sums of squares for columns 1, 2 and 3 in matrix S are, respectively

3.437 1.221 1.636

A transformation is now carried out to rotate the second and third columns of S so that the second column will account for the maximum amount of variance possible in an orthogonal space of the two vectors. When this has been done, the third column contains the remaining portion of the variance not accounted for by the second factor.

The rotated matrix D is now appended to the column vector S_1 to form B.

	I	II	III	
1.	0.504	-0.212	-0.060	
2.	0.378	-0.418	-0.550	
3.	0.781	0.153	-0.270	Items
4.	0.787	-0.123	0.151	
5.	0.354	0.780	-0.234	
6.	0.400	0.543	-0.219	
7.	0.524	0.265	-0.167	
C_1.	0.458	-0.700	-0.472	Criteria
C_2.	0.201	0.529	-0.315	
GV.	1.000	0.000	0.000	Goal Test

The respective sums of squares for the above columns are

3.437 2.003 0.854

which total to the same amount as in S but we now have each factor accounting for a decreasing amount of variance. In matrix B we have seven item vectors, two criteria vectors and a goal vector.

A test constructor must specify various parameters. The values used in this example are as follows:

SC1 = 0.200, SC2 = $\underline{b}_{il}^2$ for each respective item vector,

SC3 = 0.30, and SC4 = 45 degrees.

"Error" variance, $\underline{e}_i$, is defined as

$$e_i = h_i^2 - c_{il}^2$$

where $\underline{h}_i^2$ is the communality of the centroid vector and $\underline{c}_{il}^2$ represents the variance accounted for by the first element of the centroid row vector. Items 4 and 3 are first selected because they have the largest $\underline{b}_{il}$ values of all items available for selection

Item 3.	0.781	0.153	-0.270
Item 4.	0.787	-0.123	0.151
Sum of items	1.568	0.030	-0.119
Centroid	0.784	0.015	-0.060
Centroid variance	0.615	0.000	0.004
Communality	0.619		
h_i	0.787		
cos θ	0.784 / 0.787 = 0.996; θ = 4.465 degrees		
e_i	0.004		

The correlation of the composite vector with the goal vector is 0.784. Since the stop criteria are not applicable at this stage, we now proceed to select another item. Items 2, 5 and 6 are rejected because

in each case $h_i^2 - b_{il}^2$ is greater than b_{il}^2. Thus items 1 and 7 remain to be selected from the pool.

Sum of items previously selected	1.568	0.030	-0.119	1.568	0.030	-0.119
Item 1.	0.504	-0.212	-0.060			
Item 7.				0.524	0.265	-0.167
Sum of items	2.072	-0.182	-0.179	2.092	0.295	-0.286
Centroid	0.691	-0.061	-0.060	0.697	0.098	-0.095
h_{il}^2	0.4848			0.5044		
h_{il}	0.696			0.710		
cos θ	0.691 / 0.696 = 0.993			0.697 / 0.710 = 0.982		

Since the addition of item 1 to the composite vector will reduce the angular departure of the composite vector from the goal test vector more than item 7, item 1 is now selected. The intermediate summary data is tabulated below.

Sum of items previously selected	1.568	0.030	-0.119
Item 1	0.504	-0.212	-0.060
Sum of three items	2.072	-0.182	-0.179
Centroid	0.691	-0.061	-0.060
Centroid variance	0.477	0.004	0.004
Communality	0.485		
h_i	0.696		
cos θ	0.691 / 0.696 = 0.993; θ = 7.031 degrees		
e_i	0.008		

The correlation of the composite vector with the goal vector is 0.691.

One item remains available for selection purposes.

Sum of items previously selected	2.072	-0.182	-0.179
Item 7.	0.524	0.265	-0.167
Sum of four items	2.596	0.083	-0.346
Centroid	0.649	0.021	-0.087
Centroid variance	0.421	0.000	0.008
Communality	0.429		
h_i	0.655		
cos θ	0.649 / 0.655 = 0.991; θ = 7.798 degrees		
e_i	0.008		

The correlation of the composite vector with the goal vector is 0.649.

As there are no items remaining that meet the specified criteria, in terms of the parameters set by the user, the selection procedure is terminated.

The test constructed from four items selected in the example would have a validity of 0.649. When the composite test vector is normalized, the test validity becomes 0.991. By using the position of the centroid calculated for the constructed test, the factors can be proportionately weighted as before when using a transformation matrix. The composite test vector

$$\begin{bmatrix} 0.649 & 0.021 & -0.086 \end{bmatrix}$$

after normalization is

$$\begin{bmatrix} 0.991 & 0.032 & -0.131 \end{bmatrix}$$

Items selected from the matrix of n items

$$\begin{array}{c} \\ (1) \\ (3) \\ (4) \\ (7) \\ \\ (GV) \end{array} \begin{bmatrix} 0.504 & -0.212 & -0.060 \\ 0.781 & 0.153 & -0.270 \\ 0.787 & -0.123 & 0.151 \\ 0.524 & 0.265 & -0.167 \\ \text{---} & \text{---} & \text{---} \\ 0.991 & 0.032 & -0.131 \end{bmatrix}$$

which when postmultiplied by the normalized column vector

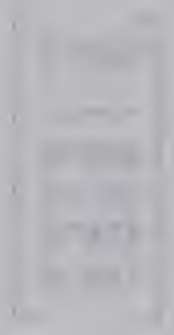
$$\begin{bmatrix} 0.991 \\ 0.032 \\ -0.131 \end{bmatrix}$$

yields the factor loadings of each selected vector on the composite test vector. Loadings on the composite test vector are

$$\begin{array}{c} (1) \\ (3) \\ (4) \\ (7) \\ \\ (GV) \end{array} \begin{bmatrix} 0.501 \\ 0.815 \\ 0.756 \\ 0.549 \\ \text{---} \\ 1.000 \end{bmatrix}$$

which, excluding the last element, has a total variance of 1.787. The proportion of variance accounted for on the composite test vector is 0.447 which is the internal consistency reliability. If the internal consistency reliability is calculated using the goal test vector as the location of the column vector, the reliability coefficient is 0.440.

It must be remembered that the reliability and validity coefficients presented above are those defined in relation to the presented analytical item-selection model. Although the notion of reliability and validity have been used, the traditional formulae have not been used because they were not appropriate.



CHAPTER VII

EVALUATION OF THE ITEM SELECTION METHOD

The proposed algorithm to be used for selecting items from a pool of items has as its foundation factor analytic theory. After a test containing many items has been administered to a group for which there are, ideally, several criterion variables available, the test items and criterion elements are factor analyzed. A rotation of the resulting orthogonal factor structure matrix is applied to provide a final solution that has simple structure properties. Each factor is then assigned a relative weight by the test user. From knowledge of the specified weights, a postulated hypothetical goal vector is constructed that precisely determines the location in the item and criterion space of a test having characteristics desired by the user. The simple structure factor matrix is then rotated to a position where factor one and the goal vector are collinear.

Initially, the two items having the highest correlations with the goal vector (the correlation is the same as the factor loading on factor one) are selected. A centroid of the composite vector is then calculated for the two item vectors. Additional items are selected and added to the composite vector which results in a shifting in coordinates of the centroid. The objective is to form a composite vector, composed of the items selected that will have nearly the same position in the item and criterion space as the goal vector. Conditions for

termination of the selection process were presented in the previous chapter.

Comparison to Other Models

Common item parameters resulting from an item analysis are (a) a correlation coefficient expressing the relationship between an item and a criterion variable or the total score of the test, and (b) an item's difficulty index from which the variance can be calculated. The information available from an item analysis is part of the data used in the proposed item selection method. Whereas the item parameters are presented independently in an item analysis, the present system provides a summary analysis utilizing factor analytic theory to define a common factor space with the co-ordinates of each item specified. Common item characteristics are defined by the use of factors. Thus, the proposed technique incorporates the data available from an item analysis and then provides an objective solution for determining which is the "best" item, "second best" item and so on. Clearly, this is a much needed procedure required for evaluating an item in relation to a test that is to be constructed.

Multiple correlation, canonical correlation and factor analysis models are based on essentially the same linear model employing classical regression equations. It was noted in reviewing multiple correlation principles that it is a superior method to use in test construction. The desirable characteristic of isolating each variable and assigning a relative weight to it for prediction purposes, common to multiple

correlation and factor analysis, has been included in the proposed algorithm. While much of classical test theory has been developed upon the assumption of unidimensionality of a test, the item-selection procedure presented here considers test multidimensionality as the general case with the unidimensional test being a special condition derived from the general model. Test reliability (internal consistency) and test validity can be considered from a factor analytic point of view as was noted earlier. Thus, the proposed algorithm, in part, takes into consideration and subsequently provides some evidence to the user of the relative estimates of test reliability and test validity coefficients.

As in scalogram analysis, reproducability is also a means of testing the accuracy of results in factor analysis. The development of a measure of homogeneity or scalability which will completely specify the bounds of interrelationships existing among all items of a scale has been extensively examined by Lingoes (1963). When selecting items by the writer's technique, the reproducability of the correlation coefficient between two variables may be considered as an indication of the amount of "true" score variance or conversely the amount of "error" variance. The variance of an item not accounted for ("error" variance) on factor one, which is collinear with the goal vector, will be spread out over the remaining orthogonal factors. However, the error variance is not considered in reproducing correlation coefficients between vectors. Thus, the calculated correlation coefficients, $\hat{r}$, are estimates of the "true" correlation coefficient, r. Differences between $\hat{r}$ and r may be

positive or negative.

Versatility of the Selection Algorithm

In keeping with the previous procedures, each item receives a weight of 1 if selected or a weight of 0 if the item is rejected. Although the simple weighting system is used, in general no restrictions are placed on the type of score that is to be assigned to each item. That is, no restrictions are imposed so that only dichotomously scored items can be used. The item score may be out of 1, 2, 3, or whatever is desired by the person scoring the test protocols.

Although the proposed method utilizes a factor analytic method, no restrictions are readily apparent as to why a method other than a principal component analysis cannot be used. The use of various types of correlation matrices appears to be only curtailed by the associated factor analytic method. Several combinations of correlation coefficients with methods of rotating factor matrices and types of factor analysis provide many possible variations for the application of the algorithm.

By describing the item and criterion vectors in a factor space, in which a constructed hypothetical goal test may be positioned in an infinitie number of locations, theoretically an infinite number of tests can be constructed from a single pool of items. The user of such a technique is primarily restricted by the location of the goal vector, the available pool of items and the restrictions or tolerance limits deemed necessary. A great deal of flexibility is available to the user which should result in an increased scope in test construction.

Item Pools

If the proposed item-selection method is used, greater effort, than presently given, and more detailed knowledge about test characteristics will probably be required to establish a pool of items. Each item should be checked for obvious flaws and modified where necessary prior to being included in an item pool. As a result of an evaluation of each item to determine whether the item should be added to the pool, the item pools should improve in quality. The increased standardization of item characteristics, which defines the universe being considered, provides information for evaluating any change in composition of the item pool. Greater summarization, than available through item analysis alone, of item properties is provided while increasing the flexibility of constructing a test.

Limitations

The complexity of test composition is magnified as the number of dimensions (factors) to be weighted increases. When working with as many as 10 factors, few items will, in the writer's experience, meet the necessary criteria for selection purposes. A matrix of 2, 3, or 4 factors is much easier to manipulate. The number of items that can be used from a pool of items to construct a test will generally increase as the complexity of the factor space is decreased.

Although the proposed item selection method is "machine dependent" because the amount of calculation involved necessitates the use of an electronic computer, no real problem is encountered since almost anyone who needs a digital computer has access to one. A point sometimes

raised in connection with computers, is that no "look" at the data and intermediate results is possible. There is nothing to prevent the computer from printing out various intermediate results, thus allowing intuition and insight to be optional, but not mandatory.

If a test constructor has access to a computer utilizing time-sharing features, computer-user interaction can facilitate immediate evaluation of a set of selected items. Thus, after evaluating the constructed test, a decision can be made to accept the selected items or to vary the desired test characteristics and subsequently select another set of items.

Insight and experience will be required in some cases. If the items in a pool defined two orthogonal clusters of items as illustrated in Figure 3 and a goal vector was then positioned midway between the clusters, the selected items would have large angular departures from the goal test vector. The size of angle between the item and specified goal test with a corresponding low correlation coefficient would indicate that the test constructor should investigate the possibility that there are no items in the pool to adequately test a particular domain. A second suggestion is that the item pool may be subdivided into two pools of items. Alternatively, two goal test vectors located at the centroid of each cluster would provide for the construction of two tests.

At first glance one limitation appears to be the necessity of always being required to have one or more criterion variables. While it is desirable to have criterion variables as elements of the correlation matrix, it is not necessary to include criteria for the purpose of factor analyzing the correlation matrix. The proposed algorithm

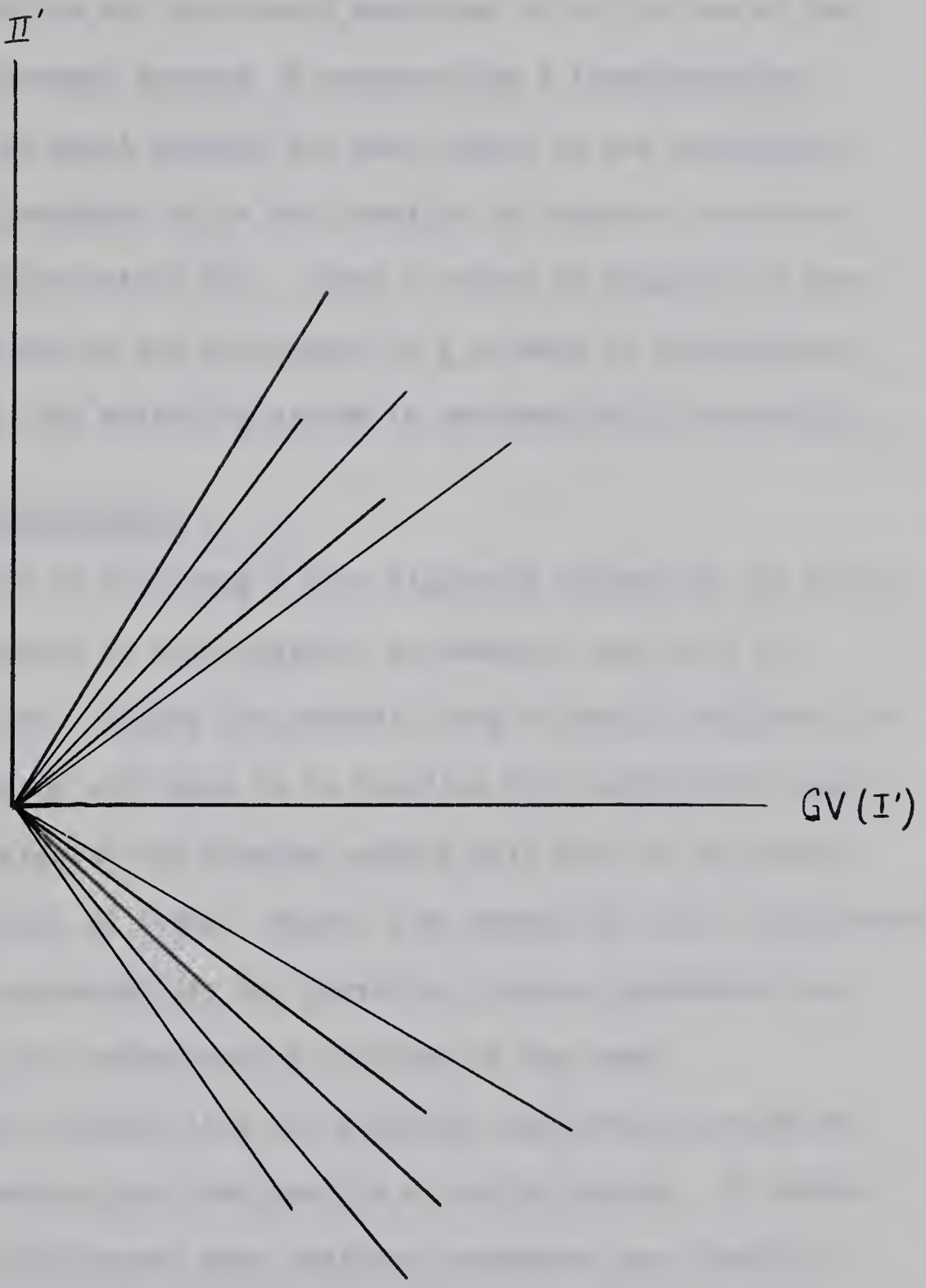


Figure 3. Orthogonal Clusters of Items

functions equally well with or without criterion components. However, if possible, criterion variables should be included.

One limitation not previously mentioned is in the use of the Gram-Schmidt orthonormal process in constructing a transformation matrix. The use of equal weights for each factor is not acceptable. With several zero weights, it is not possible to construct an orthonormal transformation matrix ($\underline{T}$). After a series of weights has been decided upon, a check on the properties of $\underline{T}$ is made by calculating $\underline{T}\underline{T}'$. If $\underline{T}\,\underline{T}' = \underline{I}$, the weighting system is mathematically acceptable.

Test Constructor Involvement

As a result of providing a more elaborate system for the selection of items compared to item analysis procedures, more will be required of the user. Values for several "stop criteria" will have to be estimated, factors will have to be labelled with appropriate names, factors must be weighted and greater concern will have to be devoted to establishing pools of items. Rather than making the test constructor's job easier, the responsibility for providing various parameters has greatly increased the understanding required of the user.

It was not intended that the proposed algorithm be presented with optimal parameters and then used in a routine manner. If better tests are to be constructed, more analytic procedures are required. However, critical decisions regarding acceptable tolerance limits and test characteristics will remain with the test constructor.

CHAPTER VIII

SUMMARY, CONCLUSIONS AND IMPLICATIONS

The problem being investigated and a proposed solution to the problem are briefly outlined. Empirical data are not, at present, available but conclusions regarding the appropriateness of the algorithm are presented. Since the present study has been concerned with a theoretical model that was not directly an extension of previous research, many theoretical and practical implications may be considered.

The Problem and a Proposed Solution

Since many items are available to construct a test, test constructors would like to know which is the "best" item, "second best" item and so on for predicting a set of criteria. The algorithm presented to solve this problem is based upon factor analytic theory. Items and criteria are factored to define a common factor space. A constructed hypothetical goal vector is defined in the factor space. The "best" k items are selected to form a composite vector that is nearly collinear with the goal vector as defined by the user.

The selection of items is dependent upon the availability of a pool of items that have been administered to a group of subjects. Parameters must be specified by the user which will result in a test being constructed according to specific desired characteristics. The item selection method provides considerable flexibility in test construction.

General Conclusions

In theory, the proposed item selection technique is directly related to a wide variety of test construction procedures such as item analysis, regression analysis and factor analysis. A logical evaluation of the algorithm for the selection technique has revealed no major difficulties regarding practical application. Because no empirical evidence is available, the present conclusions are necessarily theoretical. When evidence for the use of the proposed method and the relationship to results from other procedures are available, definite conclusions will be in order. However, as it stands, the item selection procedure appears to have merit from a theoretical viewpoint.

Implications

Although the proposed algorithm is composed of commonly used procedures, this seems to be the first time that such a practical application of an item selection method has been presented with these components. The theoretical foundation does not appear to violate the basis of measurement theory. Because of the extreme general nature of the algorithm, many problems are immediately apparent that require further research.

Theoretical. The notion of a multidimensional space, where unidimensionality is a special case of the general situation, has been considered through factor analytic theory. A variation in the procedure for factoring the items and criteria would be to factor the criteria variables and then position the item vectors in the criterion

space. Thus, the unique properties of the items in each test would not lead to variations in the stable criterion space. The implication here is that the same criteria variables would be used in comparing similar items. The use of a 'common criterion space' would provide the same marker variables for items from different tests administered to various groups.

Practical. A much needed technique has been presented. Since the item selection method is objective and at the same time flexible to the individual user, many applications should immediately be found in the routine construction of tests. Although the name "item-selection method" has been frequently used, the method need not be restricted only to the selection of items. Any variable, in the form of an element in a correlation matrix may be considered with this technique.

If items are selected from a pool of items that contains the stem and the alternatives of the question as stored information, it should be possible to prepare stencils for the final test by means of a computer and related auxiliary equipment. Such a procedure would be relatively simple to design.

Implications for Further Research. The algorithm could be used to test the effect of using various types of correlation matrices, different methods of factor analysis and variations in the procedures used for rotation of matrices to simple structure. It may be especially interesting to apply the method of alpha factor analysis (Kaiser and Caffrey, 1965) to this algorithm since the correlation coefficients in

an alpha factor analysis are corrected for attenuation during the factoring process.

Several studies are required to provide empirical evidence for the effectiveness of the item-selection method in practical test construction settings. As well as providing evidence for evaluation of the selection method, research is needed to formally and empirically compare the model to other presently available procedures such as that proposed by Wherry and Gaylord (1946).

Procedures are required that can be used to up-date item pools with additional items. The relative effects of normalizing an item vector compared to considering the communalities of each item requires investigation.

A logical examination should be carried out to examine the relevance of the size and the type of parameters used in terminating the selection procedure. Following this, a statistical evaluation should be presented to determine the relative importance of the parameters as indices.

REFERENCES

Adams, Georgie and Torgerson, T. L. Measurement and evaluation in education, psychology and guidance. New York: Holt, Rinehart and Winston, 1964.

American Educational Research Association. Technical recommendations for achievement tests. Washinton, D. C.: American Educational Research Association, 1955.

American Psychological Association. Technical recommendations for psychological tests and diagnostic techniques. Washington, D. C.: American Psychological Association, 1954.

American Psychological Association. Standards for educational and psychological tests and manuals. Washington, D. C.: American Psychological Association, 1966.

Anastasi, Anne. Psychological testing. (2nd ed.) New York: Macmillan, 1961

Astin, A. W. Criterion-centered research. Educational and Psychological Measurement, 1964, 24, 807 - 822.

Ayers, J. D. Justification of Bloom's taxonomy by factor analysis. Unpublished manuscript, University of Alberta, 1966.

Baggaley, A. R. Intermediate correlational methods. New York: Wiley, 1964.

Bloom, B. S. (ed.), Taxonomy of educational objectives, handbook I: Cognitive domain. New York: Longmans, Green, 1956.

Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelations. Psychometrika, 1946, 11, 197 - 214.

Carroll, J. B. An analytical solution for approximating simple structure in factor analysis. Psychometrika, 1953, 18, 23 - 38.

Cooley, W. W. and Lohnes, P. R. Multivariate procedures for the behavioral sciences. New York: Wiley, 1962.

Coombs, C. H. The concepts of reliability and homogeneity. Educational and Psychological Measurement, 1950, 10, 43 - 56.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297 - 334.

Cronbach, L. J. Essentials of psychological testing. (2nd ed.) New York: Harper, 1960.

Cronbach, L. J., Rajaratnam, N and Gleser, Goldine C. Theory of generalizability: A liberation of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137 - 163.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281 - 302.

Cureton, E. E. Validity, reliability, and baloney. Educational and Psychological Measurement, 1950, 10, 94 - 96.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951, pp. 621 - 694.

Davis, F. B. Item selection techniques. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951, pp. 266 - 328.

Douglas, H. and Spencer, P. Is it necessary to weight exercises in standardized tests? Journal of Educational Psychology, 1923, 14, 109 - 112.

DuBois, P. H. An introduction to psychological statistics. New York: Harper and Row, 1965.

Ebel, R. L. Measuring educational achievement. New Jersey: Prentice Hall, 1965.

Elfving, G., Sitgreaves, R., and Solomon, H. Item selection procedures for item variables with a known factor structure. Psychometrika, 1959, 24, 189 - 205.

Eysenck, H. J. Uses and limitations of factor analysis in psychological research. In Anne Anastasi (Ed.), Testing Problems in Perspective. Washington, D. C.: American Council on Education, 1966, pp. 355 - 359.

Flowers, J. F. The application of electronic digital computers to item analysis and test development. Paper read at Canadian Association of Professors of Education, 1965.

Freeman, F. S. Theory and practice of psychological testing. (revised edition) New York: Holt, 1955.

Fruchter, B. Introduction to factor analysis. Princeton: Van Nostrand, 1954.

Fruchter, B. and Jennings, E. Factor analysis. In H. Bocko (Ed.), Computer applications in the behavioral sciences. New Jersey: Prentice Hall, 1962, pp. 238 - 265.

Furst, E. Constructing evaluation instruments. New York: Longmans, Green and Company, 1958.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Glass, G. V. Alpha factor analysis of infallible variables. Psychometrika, 1966, 31, 545 - 561.

Glesser, Goldine and DuBois, P. A successive approximation method of maximizing test validity. Psychometrika, 1951, 16, 129 - 139.

Green, B. F., Jr. The computer revolution in psychometrics. Psychometrika, 1966, 31, 437 - 445.

Greene, H. A., Jorgenson, A. N. and Gerberich, J. R. Measurement and evaluation in the secondary school. New York: Longmans, Green and Company, 1954.

Guilford, J. P. Psychometric methods. (2nd ed.) New York: McGraw-Hill, 1954.

Guilford, J. P. Fundamental statistics in psychology and education. (4th ed.) New York: McGraw-Hill, 1965.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79 - 91

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950. (a)

Gulliksen, H. Intrinsic validity. American Psychologist, 1950, 5, 511 - 517. (b)

Guttman, L. Image theory for the structure of quantitative variates. Psychometrika, 1953, 18, 277 - 296.

Guttman, L. A generalized simplex for factor analysis. Psychometrika, 1955, 20, 173 - 192.

Harman, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1960.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Helmstadter, G. C. Principles of psychological measurement. New York: Appleton-Century-Crofts, 1964.

Hilgard, E. R. A test item file to accompany Hilgard's introduction to psychology 3rd edition. New York: Harcourt, Brace and World, 1962.

Hoffman, B. The tyranny of testing. New York: Crowell-Collier, 1962.

Hohn, F. E. Elementary matrix algebra. (2nd ed.) New York: Macmillan, 1964.

Horst, P. Item selection by means of a minimizing function. Psychometrika, 1936, 1, 229 - 244.

Horst, P. Optimal test length for maximum differential prediction. Psychometrika, 1956, 21, 51 - 66.

Horst, P. Relations among m sets of measures. Psychometrika, 1961, 26, 129 - 149.

Horst, P. Factory analysis of data matrices. New York: Holt, Rinehart and Winston, 1965.

Horst, P. Psychological measurement and prediction. California: Wadsworth, 1966.

Horst, P. and MacEwan, Charlotte, Optimal test length for maximum absolute prediction. Psychometrika, 1956, 21, 111 - 124.

Horst, P. and MacEwan, Charlotte, Optimal test length for multiple prediction: the general case. Psychometrika, 1957, 22, 311 - 324.

Householder, A. S. and Young, G. Matrix approximations and latent roots. American Mathematical Monthly, 1938, 45, 165 - 171.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153 - 160.

Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. Psychometrika, 1958, 23, 187 - 200.

Kaiser, H. F. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 1960, 20, 141 - 151.

Kaiser, H. F. Psychometric approaches to factor analysis. In Anne Anastasi (Ed.), Testing Problems in Perspective. Washington, D. C.: American Council on Education, 1966, pp. 360 - 368.

Kaiser, H. F. and Caffrey, J. Alpha factor analysis. Psychometrika, 1965, 30, 1 - 14.

Katzell, R. A. Symposium: The need and means of cross-validation. III. Cross-validation of item analysis. Educational and Psychological Measurement, 1951, 11, 16 - 22.

Kelly, T. L. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30, 17 - 24.

Kelly, E. L. Alternate criteria in medical education and their correlates. In Anne Anastasi (Ed.), Testing Problems in Perspective. Washington, D. C.: American Council on Education, 1966, pp. 176 - 194.

Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 95 - 101.

Layton, W. L. The relationship between the method of successive residuals and the method of exhaustion. Psychometrika, 1951, 16, 51 - 56.

Lennon, R. T. Assumptions underlying the use of content validity. Educational and Psychological Measurement, 1956, 16, 294 - 304.

Lindquist, E. F. (Ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Lindgoes, J. C. Multiple-scalogram analysis: A set-theoretic model for analyzing dichotomous items. Educational and Psychological Measurement, 1963, 23, 501 - 524.

Loevinger, Jane. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493 - 504.

Lord, F. M. A theory of test scores. Psychometric Monographs, 1952, No. 7.

Lord, F. M. Optimum level of item difficulty. Research Memo, Princeton, N. J.: Educational Testing Service, 1953.

Lord, F. M. Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 1957, 17, 501 - 521.

Lord, F. M. Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 1959, 19, 233 - 239.

Lubin, A. and Osburn, H. G. A theory of pattern analysis for the prediction of a quantitative criterion. Psychometrika, 1957, 22, 63 - 74.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122 - 131.

McNemar, Q. Psychological statistics. (3rd ed.) New York: Wiley, 1962.

Mollenkopf, W. G. Variation of the standard error of measurement. Psychometrika, 1949, 14, 189 - 229.

Mollenkopf, W. G. Predicted differences and differences between predictions. Psychometrika, 1950, 15, 409 - 417.

Mosier, C. I. Symposium: The need and means of cross-validation. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 1951, 11, 5 - 11.

Neuhaus, J. O. and Wrigley, C. The quartimax method: An analytical approach to orthogonal simple structure. British Journal of Statistical Psychology, 1954, 7_, 81 - 91.

Novick, M. R. The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 1966, 3, 1 - 18.

Nunnally, J. C., Jr. Tests and measurements - assessment and prediction. New York: McGraw-Hill, 1959.

Orleans, J. S. A test item file to accompany Cronbach's educational psychology second edition. New York: Harcourt, Brace and World, 1963.

Osburn, G. H. and Lubin, A. The use of configural analysis for the evaluation of test scoring methods. Psychometrika, 1957, 22 359 - 372.

Ray, W. S., Hundleby, J. D. and Goldstein, D. A. Test skewness and kurtosis as functions of item parameters. Psychometrika, 1962, 27. 39 - 47.

Remmers, H. H. and Gage, N. L. Educational measurement and evaluation (2nd ed.) New York: Harper, 1955.

Richardson, M. W. The relation between the difficulty and the differential validity of a test. Psychometrika, 1936, 1, 33 - 49.

Richardson, M. W. and Adkins, D. C. A rapid method of selecting test items. Journal of Educational Psychology, 1938, 29, 547 - 552.

Rozeboom, W. W. Foundations of the theory of prediction. Homewood, Ill.: Dorsey, 1966.

Ryans, D. G. Measurement and prediction of teacher effectiveness. In Anne Anastasi (Ed.), Testing Problems in Perspective. Washington, D. C.: American Council on Education, 1966, pp. 222 - 237.

Saunders, D. R. A computer program to find the best-fitting orthogonal factors for a given hypothesis. Psychometrika, 1960, 25, 199 - 205.

Stoker, H. W. and Kropp, R. P. Measurement of cognitive processes. Journal of Educational Measurement 1964, 1, 39 - 42.

Swineford, F. Note on tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 1959, 19, 241 - 242.

Terman, L. M. and Merrill, Maud A. Stanford-binet intelligence scale: manual for the third revision. Form L-M. Boston: Houghton Mifflin, 1960.

Thorndike, R. L. Personnel selection - test and measurement techniques New York: Wiley, 1949.

Thorndike, R. L. Reliability, In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951, pp. 560 - 620.

Thorndike, R. L. and Hagen, Elizabeth. Measurement and evaluation in education and psychology. New York: Wiley, 1955.

Thurstone, L. L. Multiple factor analysis. Chicago: University of Chicago, 1947.

Toops, H. A. The L-method. Psychometrika, 1941, 6, 249 - 266.

Traxler, A. E. Administering and scoring the objective test. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951, pp. 329 - 416.

Webster, H. Maximizing test validity by item selection. Psychometrika, 1956, 21, 153 - 164.

Wechsler, D. The measurement and appraisal of adult intelligence. (4th ed.) Baltimore: Williams and Wilkins, 1958.

Wherry, R. J. and Gaylord, R. H. Test selection with integral gross score weights. Psychometrika, 1946, 11, 173 - 183.

Wherry, R. J. and Winer, B. J. A method for factoring large numbers of items. Psychometrika, 1953, 18, 161 - 179.

APPENDIX A

ITEM SELECTION ALGORITHM